# AVIAN LEUKOSIS VIRUS DNA INTEGRATION: REGULATION AND SELECTION IN TUMORIGENESIS

By

Shelby Winans

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, MD
December 2017

## Abstract

A key requirement of the retroviral lifecycle is integration of the proviral genome into the host cell genome. This makes retroviral vectors uniquely suited for gene therapy applications. While murine leukemia virus (MLV) and human immunodeficiency virus-1 (HIV-1) based vectors are commonly used, we propose that avian leukosis virus (ALV) may be a safer vector. Previous gene therapy trials, while successful, had issues with integration into and activation of oncogenes leading to the formation of cancer. In this thesis, we show that ALV integration is relatively random with only slight integration site preferences in the chicken genome, potentially making it less prone to insertional mutagenesis.

To better understand how this random integration pattern is achieved, we also determine host cell factors that regulate ALV integration. Host proteins have previously been shown to target integration of MLV and HIV-1 to transcription start sites and active genes respectively. We show here that the cellular FACT (facilitates chromatin transcription) complex proteins regulate ALV integration efficiency *in vitro* and *in vivo* by binding directly to the ALV integrase protein. We show that the integration pattern of ALV *in vivo* changes significantly with varying expression levels of FACT complex. We hypothesize based on the observed direct binding of the FACT complex to the integrase protein and the observed effect *in vivo*, that the FACT complex may be acting as a bimodal tether to recruit the ALV pre-integration complex to specific genomic locations, thereby influencing both efficiency and pattern of integration.

We also explore other host cell protein candidates that selectively bound the ALV integrase protein. Interestingly, we discovered that the BET family of proteins, which are

known to regulate MLV integration, also seem to have an effect on ALV integration efficiency *in vivo*. We also see subtle effects of BET protein inhibition on integration targeting. Interestingly, BET protein inhibition in a FACT knockdown background has the largest effect on integration targeting suggesting a potential collaborative effect of the cellular host factors. Other factors explored, such as nucleolin and UBTF (upstream binding transcription factor) were found to regulate the ALV lifecycle but not directly at the level of integration.

In addition to analyzing the regulation of ALV integration, this thesis also documents how the subsequent selection of integration sites *in vivo* can be used to identify novel genes involved in tumorigenesis. Because retroviruses contain strong promoter and enhancer elements, the insertion of the proviral genome into the host cell genome can have profound effects on host gene expression. Depending on the site of integration, this can activate gene expression or promote the expression of altered gene products. *In vivo,* integration sites into genes that contribute to the regulation of proliferation, immortalization or apoptosis are selected for over time and can lead to the formation of tumors. We identified selected, or expanded, integration sites in B-cell lymphomas. This led to the identification of novel oncogenes *CTDSPL* and *CTDSPL2* as well as the putative noncoding *TAPAS* RNA.

CTDSPL and CTDSPL2 have been previously shown to regulate the phosphorylation status of the C-terminal domain of RNA polymerase II. They also play a role in regulating pRb phosphorylation and thus have been previously characterized as tumor suppressor genes. To the contrary, in our system we observe that CTDSPL and CTDSPL2 while having slight negative effects on cell proliferation, promote cell

migration and protect cells against apoptosis, attributes more characteristic of an oncogene. Truncated proteins, like those generated by ALV integration within CTDSPL and CTDSPL2 in tumors, also promote immortalization in primary cell culture. Thus, we hypothesize that integrations into these genes were selected for in tumors due to the immortalization role of the truncated protein products.

The most common expanded integration site identified in B-cell lymphomas was in the *TERT* (telomerase reverse transcriptase) promoter region. We found that these integrations were predominantly in the opposite transcriptional orientation to *TERT* and were in fact promoting the expression of a truncated form of a novel antisense long noncoding RNA, which we have named *TAPAS* (TERT antisense promoter associated) RNA. *TAPAS* RNA is conserved in most birds and we find evidence for a similar transcript in humans. We provide evidence here for a role of *TAPAS* RNA in regulating *TERT* expression in *cis* in both chickens and humans.

Thesis advisor: Dr. Karen Beemon

Thesis Committee: Dr. M. Andrew Hoyt

Dr. Robert Schleif

Dr. Greg Bowman

**Acknowledgements**

First and foremost, I would like to thank my mentor Dr. Karen Beemon. I feel fortunate to have been a part of her lab. I appreciate her support and willingness to let me independently explore and develop my scientific ideas. I am also extremely thankful to her for being a strong role model of a female professor. I would also like to thank my committee members, Dr. Schleif, Dr. Hoyt and Dr. Bowman for their scientific input and support.

I am also extremely thankful to the members of the Beemon lab, in particular, Yingying, who has been a lab mom to all of us. In addition to being a wonderful technician and extremely helpful with experiments, she has been a great friend over the years listening to all of my complaints and always cheering me on. I also am incredibly grateful to my best labmate, Sunny. We joined the lab at the same time and you've been there every step of the way always supporting me – from GBO preparation to progress reports to interviews. Thank you for listening to all my talks a million times, discussing experiments with me and putting up with my competitive spirit in lab and all of my bad days. You truly made me look forward to coming in to lab every day and I don't know what I would have done without you over the years.

To my wonderful friends, near and far, who keep me grounded and always lift my spirits, I am incredibly appreciative. To my friends from home, your pride and belief in me over the years has been invaluable. A special mention must go to Maddy, who has been my other half for years and has always been there for moral support. I am thankful to my graduate school friends, particularly Sabrina and Riti, for helping me keep things in perspective and keep life fun over the past few years.

To my parents, I am eternally indebted. They have been the most supportive and proud parents a person could ask for. Despite not knowing what I'm talking about most of the time, they were always there to listen. I am particularly grateful to my mother for instilling in me a love of learning that has surely carried me through grad school. Being a first generation college student was hard for us all and I know that letting me go and explore the big cities has been scary for you but your support has made all the difference. I would not be who I am today without you. I am also thankful to my brother, whose support and competitive spirit, has pushed me to become a better version of myself in many ways.

I must also thank Marty, my high school science teacher, who I most certainly would not be where I am today without. He recognized my potential early on and has mentored me for more than a decade. He pushed me to be better, reach further and achieve more. I am appreciative to him for the many, many conversations about my goals and my career. He has put up with my whining for years and has truly helped me to keep things in perspective.

Lastly, I would like to thank Max. His support throughout college and grad school has been invaluable. From making sure I was fed and caffeinated throughout the early mornings and late nights in lab to supporting me mentally through GBOs and thesis writing. He has taken care of me physically and mentally for many years now and I don't know what I would have done without him.

Thanks to everyone who has made this experience possible and enjoyable!

**Table of Contents**

**List of Tables:**

**List of Figures**

## Abbreviations

| | |
|---|---|
| **ALV** | **Avian leukosis virus** |
| **MLV** | **Murine leukemia virus** |
| **HIV-1** | **Human immunodeficiency virus type-1** |
| **HTLV** | **Human T-lymphotropic virus** |
| **MMTV** | **Mouse mammary tumor virus** |
| **PFV** | **Prototype foamy virus** |
| **FACT** | **Facilitates chromatin transcription** |
| **SSRP1** | **Structure specific recognition protein 1** |
| **SPT16** | **Suppressor of Ty protein 16** |
| **IN** | **Integrase** |
| **RT** | **Reverse transcriptase** |
| **BET** | **Bromodomain and extra-terminal domain** |
| **UBTF** | **Upstream binding transcription factor** |
| **NCL** | **Nucleolin** |
| **TERT** | **Telomerase reverse transcriptase** |
| **TAPAS RNA** | **TERT antisense promoter-associated RNA** |
| **CTD** | **C-terminal domain** |
| **CTDSPL** | **CTD small phosphatase-like protein** |
| **CTDSPL2** | **CTD small phosphatase-like protein 2** |
| **LEDGF** | **Lens epithelial derived growth factor** |
| **CPSF6** | **Cleavage and polyadenylation specific factor 6** |
| **CEF** | **Chick embryo fibroblasts** |
| **FPKM** | **Fragments per kilobase per million reads** |
| **LTR** | **Long terminal repeat** |
| **TSS** | **Transcription start site** |
| **RIG** | **Recurrent integration gene** |
| **CIS** | **Common integration site** |

# Chapter 1 – Introduction to Retroviruses and Integration

## 1.1 Research Objectives

My thesis work has been divided into two major areas of study concerning avian leukosis virus (ALV) integration. The first goal of my work was to gain a better understanding of the integration pattern of ALV and how it differs from other retroviruses. In addition, I attempted to understand the cellular host factors that regulate integration and determine ALV integration site selection *in vivo*.

The goal of my second project was to identify novel players in oncogenesis using ALV as an insertional mutagenesis tool. Making use of high throughput sequencing of multiple ALV-induced tumors we identify genes that harbor clonally expanded, recurrent integrations and subsequently characterize the role of these genes in oncogenesis.

## 1.2 Retroviral classification

Avian leukosis virus (ALV) belongs to the retroviridae family, which consists of viruses with a single strand RNA genome that utilize a DNA intermediate to replicate. The family consists of seven genera, alpha- through epsilon-retrovirus, lenti- and spumavirus. Retroviruses are ascribed to a genera based on sequence relatedness of a designated portion of the reverse transcriptase open reading frame, found to be the most conserved region amongst retroviral family members. Other criteria to classify retroviruses include virion core morphology as well as the presence of accessory open reading frames. For instance, alpharetroviruses and gammaretroviruses have been classified as simple retroviruses due to the presence of only the common core retroviral genes – *gag, pol* and *env*, whereas HIV-1 has been classified as a "complex" retrovirus due to the presence of many overlapping ORFs and accessory proteins (Coffin, et al. 1997).

### 1.3 Genome structure and organization

As mentioned, ALV is a simple retrovirus with just the three open reading frames – *gag, pol* and *env* – which are common to all retroviruses. The *gag* gene of ALV encodes the structural proteins matrix (MA), capsid (CA) and nucleocapsid (NC) as well as the protease (PR) enzyme. The matrix protein, encoded by the N-terminal portion of *gag*, associates with the lipid membrane of the cell and is important in virion assembly (Hamard-Peron and Muriaux, 2011). The capsid protein forms the structural protein core that encapsidates the RNA genome. Nucleocapsid is a small (60-90 amino acids) basic protein that coats the viral RNA genome. NC plays important roles in facilitating the annealing of tRNA to the primer binding site as well as in formation of the dimeric RNA genome (Feng et al., 1996). NC mutations have also been reported to block RNA packaging implicating the protein in virion assembly as well (Darlix et al., 2014). The protease protein is required for proteolytic cleavage of the viral polyprotein during virion maturation (Coffin et al., 1997).

*pol* encodes the reverse transcriptase (RT) and integrase (IN) enzymes. The RT protein possesses reverse transcriptase catalytic activity as well as RNase H activity, both of which are required for catalyzing reverse transcription of the viral RNA genome to a double stranded DNA genome. The coding sequence of RT is one of the most conserved sequences amongst retroviruses. However, despite conservation, the active subunit structure differs between genera. For instance, in ALV, RT functions as a heterodimer with one dimer consisting of only the polymerase and RNase H domains and the larger subunit consisting of an additional fused integrase (IN) domain. HIV-1 RT likewise functions as a heterodimer but with one subunit possessing only polymerase activity and

one subunit possessing both polymerase and RNase H catalytic activity. MLV RT on the other hand is able to function as a monomer.

The *env* mRNA forms from the splicing of the viral leader sequence, in the extreme N terminus of gag (+18 nt), to a splice acceptor site in *env*. Env encodes the surface glycoprotein (SU) and transmembrane (TM) protein of the virion. Because these glycoproteins bind cellular surface proteins, they are responsible for determining the host cell range of the retroviruses (Coffin et al., 1997).

Lastly, the viral genome is flanked by long terminal repeats (LTRs) that contain strong promoter and enhancer elements to drive the expression of viral genes (Coffin et al., 1997). A portion of the LTR sequences are also required for integration.

**1.4 Retroviral life cycle**

When a virion approaches a susceptible host cell, binding of the viral surface glycoprotein to an appropriate host cell receptor triggers endosomal uptake of the virion. Upon acidification of the endosome, fusion of the viral envelope and host cell membrane occurs leading to the release of the viral capsid into the cellular cytoplasm (Barnard and Young, 2003). Once in the cytoplasm reverse transcription of the viral RNA by the viral RT protein begins, converting the RNA genome into a double stranded DNA copy utilizing a virally packaged tRNA primer and cytoplasmic dNTPs (Coffin et al., 1997). In the case of ALV specifically, reverse transcription is completed in the nucleus (Werner et al., 2002).

The preintegration complex (PIC) forms near the end of reverse transcription in the cytoplasm and consists of the critical viral integrase (IN) protein as well as other viral and cellular proteins required for efficient integration *in vivo*. The components of the PIC

are poorly understood due to the low abundance of the PIC during infection (i.e. one copy per cell) (Engelman and Cherepanov, 2017).

Depending on the family of retrovirus, breakdown of the nuclear membrane may be required for the retroviral PIC to access the host genomic DNA. This is true of most simple retroviruses, such as murine leukemia virus (MLV). However, more complex retroviruses, such as HIV-1, are able to infect non-dividing cells. The PIC of HIV-1 is known minimally to contain MA protein as well as an HIV-1 specific accessory protein, Vpr, both of which possess nuclear localization signals (NLS) and thus are believed to contribute to the active transport of the PIC into the nucleus (Gallay et al., 1995; Heinzinger et al., 1994). More recently, the HIV-1 integrase protein was found to also facilitate nuclear import both by possessing an NLS of its own and recruiting MA to the PIC (Gallay et al., 1997). Further, differences in time of capsid uncoating of the viral genome may play a role in viral ability to enter the nucleus of nondividing cells. HIV-1 has been shown to uncoat early in the life cycle, at the time of reverse transcription, whereas MLV capsid remains associated with the viral genome until at least nuclear entry (Fassati, 2006; Yamashita and Emerman, 2006).

ALV exhibits an intermediate phenotype between that of MLV and HIV-1 (Hatziioannou and Goff, 2001). ALV is able to infect nondividing cells significantly better than MLV, but still with a decreased efficiency as compared to HIV-1. Similar to HIV-1, the IN of ALV has been found to possess an NLS, which could explain the ability of ALV to transduce nondividing cells. The discrepancy between the efficiency of HIV-1 and ALV import could be due to the presence of multiple NLS in PIC components of HIV-1 and only one identified NLS in the ALV PIC. Once in the nucleus, integration of

the proviral DNA genome into the host cell genome occurs catalyzed by the virally encoded integrase protein. The early steps of the life cycle described are depicted in Figure 1.1.

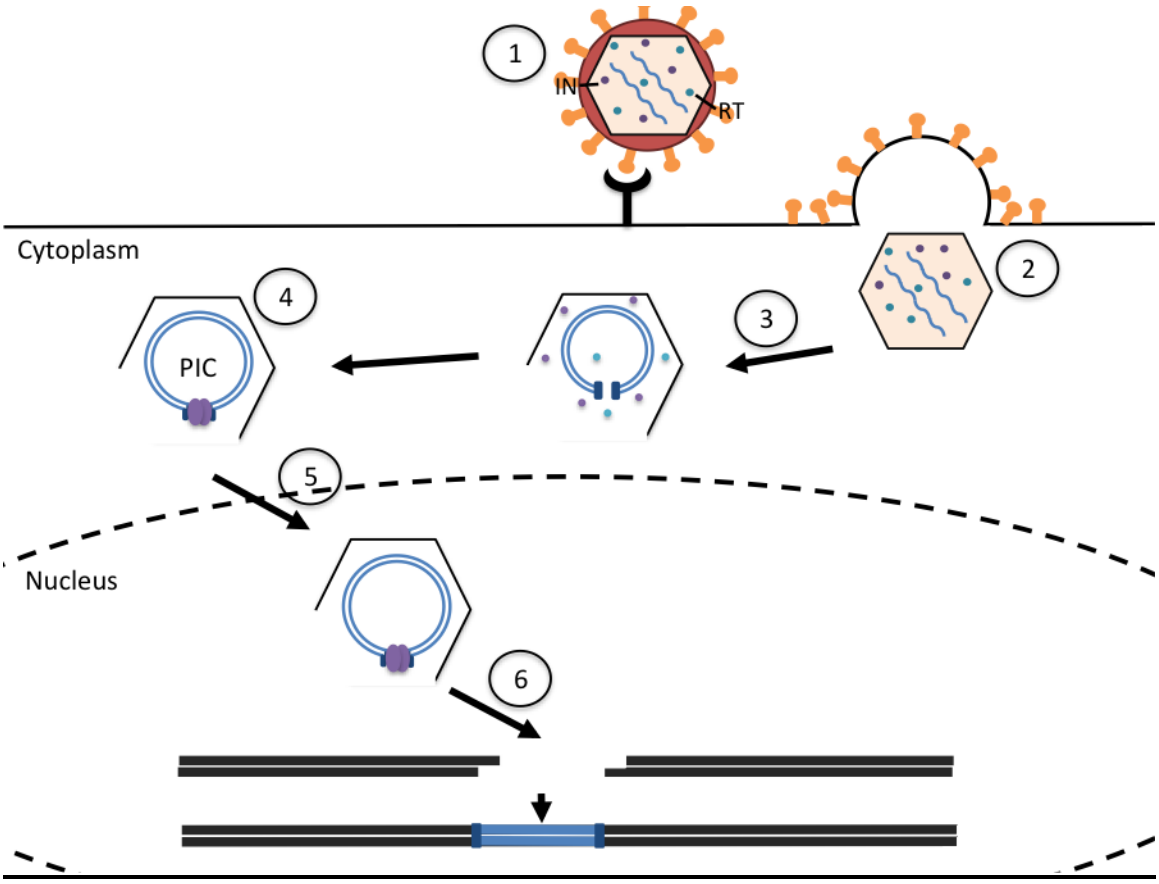



**Figure 1.1: Early steps of retroviral life cycle**. (1) Surface glycoproteins of the virion bind to a compatible receptor on the surface of the host cell. (2) This binding triggers either endosomal uptake of the virion or membrane fusion, releasing the virion capsid into the cytoplasm. (3) Reverse transcription by the virally packaged RT protein converts a dimeric single strand RNA genome into a dsDNA copy. The dsDNA copy of the genome remains in complex with RT, IN, CA and other viral proteins to form the pre-integration complex (PIC) (4). (5) The PIC gains access to the nucleus of the host cell either by active import or nuclear membrane breakdown. (6) Integration into the host cell genome by the virally packaged IN protein establishes a provirus.

After integration, the viral genome is transcribed and translated as a cellular gene by RNA polymerase II and ribosomes respectively. Gag and Pol are translated as a polyprotein whereas Env gets translated from a subgenomic mRNA. However, in ALV the *gag* and *pol* genes are in different open reading frames and thus, to get translation of the full Gag-Pol polyprotein, a programmed ribosomal frameshift is required. The frameshifting event occurs in only a small proportion of total translation events thus ensuring a high Gag to Gag-Pol ratio (Shu-Yun Le et al., 1991). This has been shown to be important for efficient viral replication (Dinman et al., 2002).

After translation of viral proteins, the process of assembly and particle budding is coordinated largely by the Gag protein. The N-terminal portion of Gag is required for proper targeting of both Gag and Gag-Pol proteins to the cellular plasma membrane. The NC portion of the Gag polyprotein is thought to be important for selective packaging of the viral RNA genome (Lu et al., 2011). Further, a number of cellular RNAs get packaged in the virion as well. However, the specific tRNA required to prime reverse transcription is enriched amongst packaged cellular RNAs. Data shows that the RT protein mediates the selective packaging of the appropriate tRNA primer (Cen et al., 2002).

The process by which Env proteins are specifically directed to the site of virion assembly is largely unknown. It is thought that Gag may also play a role in this process via the MA protein, which has been shown to bind the TM protein of Env. Budding of the assembled virion is dependent on Gag alone. After budding, the proteolytic processing of the Gag and Gag-Pol precursor proteins within the virion by viral protease (PR) is absolutely required for the maturation of infectious viral particles.

### 1.5 Importance of the viral integrase protein

A unique feature of retroviruses is their ability to integrate their cDNA genome into the host cell genome. Once integrated into the host cell genome, the provirus persists indefinitely and is passed to daughter cells via cell division. Integration is an essential step in the retroviral life cycle and requires only the integrase protein and the ends of the long terminal repeats that flank the viral genome (Donehower and Varmus, 1984; Panganiban and Temin, 1984). Aside from essential functions in integration, integrase also plays a pivotal role in various stages of the retroviral life cycle as evidenced by the array of phenotypes exhibited by integrase mutants (Engelman, 1999).

Integrase mutants have been classified into two categories based on the exhibited phenotype. Class I mutants are defined as mutations that directly affect the catalytic ability of IN. These include mutations of the IN active site. Direct effects on integrase catalytic activity are assessed by the presence of 2-LTR circles. Once thought to be the precursor to integration, 2-LTR circles have since been found to be a dead end product of failed retroviral integration (Bukrinsky et al., 1993; Panganiban and Temin, 1984). 2-LTR circles form from unintegrated viral genomes in the nucleus through the host non-homologous end joining pathway (Li et al., 2001). Thus, mutants that directly inhibit only integration will manifest with increased abundance of 2-LTR circles due to the increased presence of unintegrated viral genomes in the nucleus. Class II IN mutants on the other hand display pleiotropic effects with defects observed in reverse transcription, nuclear import and virion maturation (Charmetant et al., 2011; van Gent et al., 1993; Kessl et al., 2016; Lu et al., 2005).

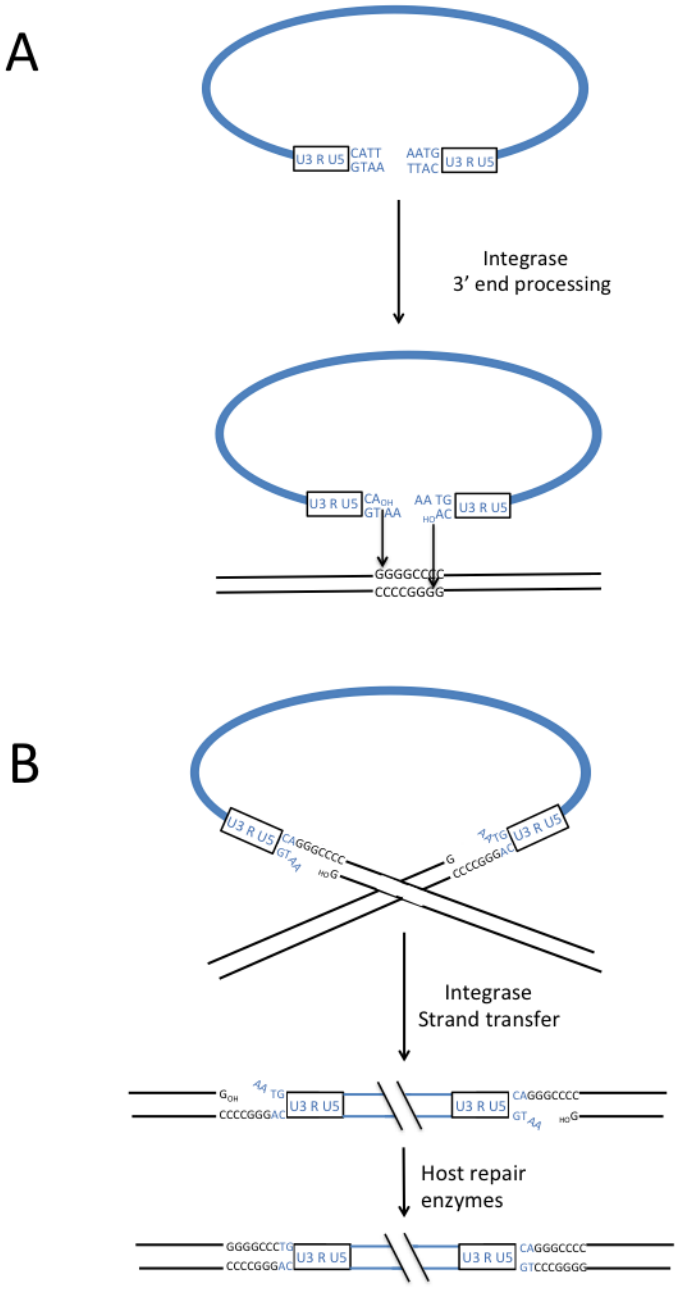
### 1.6 Molecular mechanism of retroviral integration

The integrase protein, along with several other viral and cellular proteins, combines to form the PIC in the cytoplasm of the host cell. These complexes can be isolated from infected cells and have been shown to be capable of mediating integration *in vitro* (Bowerman et al., 1989; Brown et al., 1987). The composition of the PIC varies between retroviral genera but is known to consist minimally of viral DNA, and the viral proteins - NC, MA, RT, IN, and varying levels of CA.

*In vitro* the viral IN protein alone is capable of catalyzing proviral integration into the host cell genome and does so by a two-step mechanism (Katz et al., 1990; Nowotny, 2009). First, IN cleaves the terminal 2 bases from the 3' end of the proviral genome leaving exposed 3'-OH groups immediately following an invariant CA dinucleotide (Figure 1.2A). The 3' end processing step is required for successful integration and has also been shown to be coupled to the formation of a stable integrase-DNA complex (Vink et al., 1994). Following viral DNA end processing, the exposed 3'-OH groups attack staggered 5' phosphoryl bonds of opposite strands of target host DNA (Bushman et al., 1990; Engelman et al., 1991). Both end processing and strand transfer reactions are mediated by SN2 transesterification reactions (Engelman et al., 1991).

The strand transfer reaction leaves behind single stranded gaps on either strand as well as a 2-bp overhang of the viral DNA (Figure 1.2B). Host cell repair mechanisms are thought to fill in the gaps that flank the site of integration. In the case of ALV, integration, this repair primarily generates a 6-bp repeat sequence flanking the site of integration (Hishinuma et al., 1981). The exact host cell machinery that mediates repair of the viral-host genome junction is unknown.

**Figure 1.2: Mechanism of retroviral DNA integration into the host cell genome**. (A) 3' end processing of viral DNA ends by virally encoded integrase protein. Two nucleotides at the 3' ends of the viral genome, shown in blue, are resected leaving behind an invariant CA dinucleotide with an exposed 3'-OH group. (B) Strand transfer reaction. Via an SN2 transesterification reaction the exposed 3'-OH groups attack staggered 5' phosphoryl groups of the target DNA sequence. Target DNA cleavage and subsequent strand transfer leaves behind a single strand gap and a 2-nt viral sequence overhang that is repaired by host repair machinery to form an intact provirus.


A
U3 R U5
CATT
GTAA
AATG
TTAC
U3 R U5
Integrase
3' end processing
U3 R U5
U3 R U5
B
U3 R U5
U3 R U5
Integrase
Strand transfer
U3 R U5
U3 R U5
Host repair
enzymes
U3 R U5
U3 R U5

Integration is analogous to DNA transposition, and retroviral integrase proteins belong to a superfamily of DD(E/D) transposases (Nowotny et al., 2009). IN functions as a multimer, referred to as the intasome, and the extent of multimerization varies by retroviral genus (Engelman and Cherepanov, 2017). The canonically studied model for intasome assembly and mechanism of action is the prototype foamy virus (PFV). PFV has been found to possess the lowest order IN-to-viral DNA ratio and thus this minimalist assembly serves as a straightforward model for understanding intasome architecture and function. All retroviral IN proteins consist of three conserved domains. The N-terminal domain (NTD), catalytic core domain which harbors the DDE catalytic triad and RNase H fold (CCD), and the C-terminal domain (Engelman and Cherepanov, 2017). The PFV intasome consists of four IN proteins with an IN dimer binding each of the viral DNA ends (Hare et al., 2010). The inner IN molecules provide the active sites to catalyze both the 3' processing and strand transfer reactions. The C-terminal domains of the associated IN molecules are required to engage the target DNA (Cherepanov et al., 2011; Maertens et al., 2010). Interestingly, all domains of the outer IN subunits are dispensable for catalytic activity.

The intasome architecture is not strictly conserved and a surprising amount of variability in extent of multimerization exists amongst intasomes of different retroviruses. The functional integrase multimer of alpharetroviruses (including ALV) and betaretroviruses (i.e. MMTV) was found to be an octamer (Ballandras-Colas et al., 2016; Yin et al., 2016). The HIV-1 intasome is heterogeneous. There have been reports of a functional tetrameric intasome as well as higher order assemblies with up to 16 integrase proteins (Krishnan and Engelman, 2012). Regardless of the number of integrase proteins

contained in the intasome, the functional organization is tetrameric – for PFV, IN acts as a tetramer of monomeric IN proteins; for ALV and MMTV, IN acts as a tetramer of dimeric proteins and so on. The mechanism of intasome assembly is not well understood, and while the structural features of the intasome have been uncovered for various retroviruses, the functional consequences of higher order multimerization remains unclear.

**1.7 Proviral integration is regulated by host cell factors**

While integrase alone can catalyze integration of the proviral genome, host cell factors have been found to increase efficiency of the reaction. When PICs isolated from infected cells are subjected to high salt concentration, they lose a substantial amount of catalytic activity *in vitro* (Chen and Engelman, 1998; Farnet and Bushman, 1997; Lee and Craigie, 1994). However, when extracts from uninfected cells are added back to the reaction, catalytic integrase activity is restored. Such an assay has been used to identify host cell factors that might be important for efficient integration activity.

This method identified a number of factors that were able to reconstitute integration activity *in vitro*, including BAF (barrier to autointegration factor). *In vitro* MLV PICs often use their own DNA as an integration target, a process called autointegration. BAF is a cellular protein responsible for condensing DNA, and was found to block autointegration of the proviral DNA (Lee and Craigie, 1998; Suzuki and Craigie, 2002). Moreover, BAF was found to more generally inhibit integrase activity in *in vitro* assays. Making use of the same assay with HIV-1 PICs, HMGA was identified as a novel factor that stimulates integrase activity (Farnet and Bushman, 1997). However, the importance of these host cell factors *in vivo* has been contentious.

Another method used to discover host cell factors that regulate integrase activity was to identify proteins that selectively bind the integrase protein. This was initially done making use of a yeast two-hybrid assay (Kalpana et al., 1994). This method uncovered the INI1 (integrase interactor 1) factor that has been demonstrated to stimulate HIV-1 integrase activity *in vitro*. INI1 is a member of the SWI/SNF chromatin remodeling complex. It is believed to stimulate integration in vivo by stabilizing the HIV-1 PIC and maintaining integrase in a stable, active conformation (Suzuki et al., 2012).

**1.8 Different families of retroviruses have distinct integration site preferences**

It was initially believed that integration sites were mandated by chromosome availability and that open chromatin would serve as a better integration substrate. However, all studied retroviruses exhibit distinct, non-random integration patterns (Demeulemeester et al., 2015; Derse et al., 2007; Mitchell et al., 2004; Schröder et al., 2002; Wu et al., 2003).

Retroviruses show weak sequence preferences *in vitro* and *in vivo* but this is thought to play only a minor role in integration site selection (Pryciak and Varmus 1992). Further, retroviruses have differential preferences for integrating into nucleosomal DNA (Benleulmi et al., 2015; Pryciak and Varmus, 1992). HIV-1 and ALV preferentially integrate into nucleosome poor regions, while MLV and PFV prefer to integrate into stable and compact chromatin (Benleulmi et al., 2015).

For those retroviruses that do prefer a nucleosomal target, the DNA at SHL (superhelical location) +3.5 or SHL -3.5, where the major groove of the DNA is most exposed, is the primary target for integration, suggesting that DNA distortion might favor integration in these cases (Maskell 2015). This is supported by other studies that found

that DNA distortion outside of the nucleosome also supports integration (Bor et al., 1995; Müller and Varmus, 1994). While nucleosomes and other proteins that distort DNA may favor integration, proteins bound to the DNA can also inhibit integration by steric hindrance (Pryciak and Varmus, 1992).

Despite some cursory similarities, the integration patterns of different retroviruses vary significantly (Table 1.1). For instance, HIV-1 has a very strong preference for integrating into gene regions with as much as 75% of integrations falling into gene bodies (Schröder et al., 2002). Integration of HIV-1 specifically favors expressed genes as well as highly spliced genes, but integration across the gene body is fairly uniform (Mitchell et al., 2004; Singh et al., 2015).

Murine leukemia virus (MLV) on the other hand has a notable bias for integrating in the proximity of enhancer regions, CpG islands and transcription start sites (TSS) (Lafave et al., 2014; Mitchell et al., 2004; Wu et al., 2003). MLV was initially characterized to favor TSS and CpG islands with as much as 25% of integrations falling within 5 kb of TSS and 15% of integrations occurring within 1 kb of CpG islands. However, recent work suggests that the vast majority of MLV integration sites are located near nucleosomes with the H3K4me1 epigenetic mark, a hallmark of enhancer regions (De Ravin et al., 2014).

ALV possesses one of the most random integration patterns observed with only slight preferences for transcribed genes (Barr et al., 2005; Mitchell et al., 2004; Withers-Ward et al., 1994). However, these previous studies used relatively little data to draw conclusions on ALV integration pattern. The studies were limited both by the number of integration sites they were able to sequence and also by the limited availability of an

annotated chicken genome and appropriate bioinformatics tools to correlate integrations with various genomic features. In Chapter 2, this thesis focuses on better characterizing the integration pattern of ALV in a high throughput manner.

Previous studies also largely focused on characterizing the integration pattern of ALV in human cells (Narezkina et al., 2004). In this work we characterize ALV integration in the natural host species. We specifically chose to work with DT40 cells, a chicken B-cell line derived from an ALV-induced bursal lymphoma (Winding et al. 2001). Since ALV infection *in vivo* most commonly induces B-cell lymphomas, DT40 cells are a highly relevant cell line in which to analyze natural integration patterns of ALV.

| Genomic feature | Random | HIV-1 | MLV | ALV |
|---|---|---|---|---|
| Within 5 kb of TSS | 5% | 6.9% | 26.1% | 8.4% |
| Within 1 kb of CpG island | 1% | 0.2% | 11.8% | 3.1% |
| Within genes | 33% | 77.9% | 44.3% | 42.0% |

**Table 1.1: Summary of integration site preferences of studied retroviral genera.** Shown are the frequencies with which integration is observed near transcription start sites (TSS), CpG islands and within gene bodies for HIV-1, MLV and ALV. For reference, the frequency with which you would expect integrations into these features if integration were entirely random is also shown. Data was adapted from Lewinski et al., 2006; Narezkina et al., 2004.

### 1.9 Host factors regulate integration site selection

These unique integration site preferences of retroviruses have been shown to be the result of binding of the integrase proteins by distinct host cell factors. Retroelements have long been known to make use of cellular binding partners to direct integration into very specific locations throughout the host genome. In a seminal study, the yeast Ty5 retroelement was shown to integrate preferentially into the heterochromatic telomere region and silent mating loci. This targeting activity was narrowed down to a 6 amino acid in the targeting domain of the Ty5 element integrase that was found to interact with the Sir4p protein, a structural component of silent chromatin (Gai and Voytas, 1998; Xie et al., 2001; Zou et al., 1996). Likewise, interaction between the TFIIIB complex and Ty3 integrase targets integration of the element to within 2 nucleotides of RNA Pol III transcribed genes (Bridier-Nahmias et al., 2015). In both cases, the host cell factors bound both their natural DNA binding site as well as the integrase protein, serving to bring the two into close proximity, and thereby facilitate integration at these locations.

Host cell factors have likewise been found to be responsible for the observed integration site selection biases of different retroviral genera. These factors are similarly believed to act largely as bimodal tethers linking the PIC to distinct locations throughout the genome (Kvaratskhelia et al., 2014). The first such targeting factor identified was LEDGF (lens epithelial derived growth factor). LEDGF is a general transcriptional co-activator that is found predominantly in gene regions. LEDGF was first found to stimulate HIV-1integrase catalytic activity *in vitro,* suggesting it may play a role in facilitating HIV-1 integration (Maertens et al., 2003). Upon knockdown of cellular LEDGF, HIV-1 integration efficiency is significantly decreased supporting this

hypothesis. Further, depletion of LEDGF caused a significant decrease in HIV-1 integration in transcriptional units *in vivo* indicating a notable role for this cellular factor in targeting HIV-1 integration to genes (Ciuffi et al., 2005; Llano et al., 2006; Maertens et al., 2003). However, even in the absence of LEDGF, integration of HIV-1 into gene regions remained elevated above random. A second factor, CPSF6 (cleavage and polyadenylation specific factor 6) has also been shown to play a role in targeting HIV-1 integration to gene-tropic regions of the genome (Sowd et al., 2016). Interestingly, CPSF6 was found to bind the HIV-1 CA protein rather than the IN protein indicating that conventional IN focused studies may be limiting.

Host cell factors that regulate integration of other retroviral families, such as the gammaretroviruses, have also been recently discovered. The BET (bromodomain and extraterminal) family of proteins was found to play a significant role in promoting and targeting MLV integration (De Rijck et al., 2013; Sharma et al., 2013). The BET proteins were identified as binding partners of MLV IN using an affinity chromatography approach coupled with mass spectrometry. The BET proteins are a class of bromodomain containing proteins that bind acetylated lysines of H3 and H4 tails, a common epigenetic modification found at transcription start sites (Florence and Faller, 2001). Inhibiting the BET proteins ability to bind chromatin significantly decreased MLV integration efficiency as well as targeting of integration to transcription start sites, suggesting that the BET proteins likely also act as bimodal tethers linking the PIC to specific genomic locations (Sharma et al., 2013).

The cellular serine/threonine protein phosphatase 2A (PP2A) was recently identified as a selective binding partner of deltaretroviral (i.e. HTLV-1) IN protein. It was shown to

bind IN directly and stimulate in vitro integration activity (Maertens, 2016). However, PP2A does not bind chromatin and thus does not likely act as a bimodal tether as hypothesized for LEDGF and BET. The importance of PP2A in regulating deltaretroviral integration *in vivo* also remains unclear.

Due to the pervasive utilization of host cell factors to regulate integration site selection amongst retrotransposons and retroviruses alike, it seems likely that ALV utilizes a similar mechanism to mediate integration pattern. However, the factors responsible for the more random integration pattern of ALV were previously unknown. Chapters 3 and 4 of this thesis elucidate the host cell factors that aid ALV integration efficiency as well as target integration sites.

**1.10 Consequences of integration**

That integration of the viral genome is an obligate part of the retroviral life cycle makes retroviruses uniquely well suited for use in gene therapy. The ability to integrate into the host genome is an attribute ideal for stable, long-term transgene expression. The first human gene therapy trials over 25 years ago attempted to treat severe combined immunodeficiency syndrome (SCID) using a retroviral vector system to deliver a wild type copy of the causative mutated gene (Dornburg, 2003; Rans and England, 2009). However, throughout these early studies, patients remained on drug therapy and thus, the benefit of the gene therapy treatment alone was ambiguous. In 2000, two landmark clinical SCID trials in France and England were incredibly successful, with over 85% of patients being cured (Cavazzana-Calvo et al., 2000; Hacein-Bey-Abina et al., 2002). It seemed that the promise of gene therapy had finally

materialized. However, following treatment, nearly a quarter of the patients in the trials developed T-cell leukemia (Hacein-Bey-Abina et al., 2008; Howe et al., 2008).

Retroviruses had been known to be associated with cancer in the past, most notably, Rous Sarcoma Virus, or RSV in chickens (Rous, 1910). Research on RSV, a derivative of ALV, had revealed that with establishment of an integrated viral genome, or provirus, retroviral elements are produced that flank the genome. These so-called long terminal repeats (LTRs) contain strong enhancer and promoter elements that attract host proteins to drive the expression of viral proteins (Gowda et al., 1988; Laimins et al., 1984). If viral integrations happen near host genes, the strong viral regulatory elements can also affect the expression of these genes (Hayward et al., 1981). In addition to the positive regulatory elements, proviral integration also inserts splice donor and acceptor sites that can lead to production of a host-viral fusion protein with altered regulation or function (Neel et al., 1981).

Sequencing of leukemic cells from afflicted gene therapy patients identified retroviral vector integrations near proto-oncogenes, most notably *LMO2* (McCormack and Rabbitts, 2004). This highlights one of the most serious risks with using retroviral vectors – that integration into the host genome is uncontrolled.

Human gene therapy trials to date have largely focused on the use of MLV and HIV-1 based vectors for gene delivery. While both viral vectors possess broad tropism and good expression in human cells, they also possess harmful integration patterns with strong preferences for integrating in and around regulatory regions and active gene regions respectively. Thus, there is a significant risk of deleterious insertional mutagenesis, a problem that manifested in early gene therapy trials. Avian leukosis virus

(ALV) however, may make a safer alternative due to its relatively random integration pattern. Therefore, understanding the mechanism by which retroviral integration is naturally targeted to specific locations throughout the genome in general may aid in the development of safer retroviral gene therapy vectors.

While insertional mutagenesis by integration of the retroviral genome into the host cell genome poses a substantial risk in human gene therapy, this process can also be used as a tool to identify novel gene players involved in tumorigenesis.

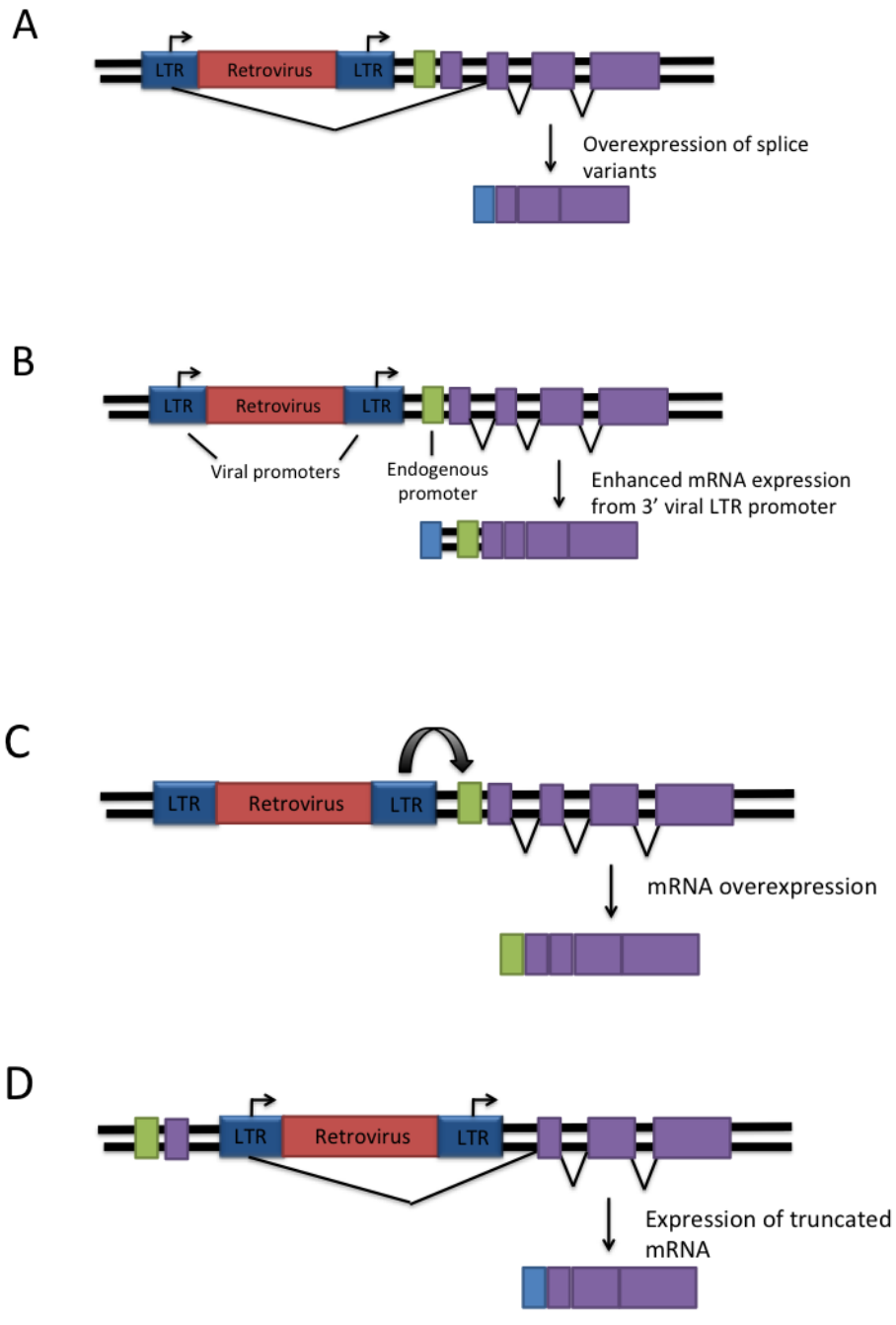
**1.11 ALV as an insertional mutagenesis tool**

ALV induces tumors in chickens by insertional mutagenesis (Beemon and Rosenberg, 2012). ALV typically induces B-cell lymphomas, but can also induce erythroblastomas, hemangiomas, and myeloid tumors (Beemon and Rosenberg, 2012; Justice and Beemon, 2013; Justice et al., 2015a). As mentioned previously, proviral integration can lead to the misregulation of nearby host genes or the production of altered viral fusion proteins which can subsequently lead to tumorigenesis. Viral integrations can perturb host genes by various mechanisms (Figure 1.3). The first such described mechanism was promoter insertion. Both the 5' and 3' LTRs contain strong promoter and enhancer elements and if the virus integrates in the same transcriptional orientation as a nearby gene, the strong viral promoter from either the 5' or 3' LTR can be utilized to drive high levels of expression of the nearby gene, If the 5' LTR is used, the viral splice donor often splices into downstream exons of adjacent genes (Figure 1.3A). If the 3' LTR is used instead, transcription proceeds through the viral poly(A) site in a process called readthrough (Figure 1B; Hayward et al., 1981). This can lead to the overexpression of full length transcripts of adjacent genes. Further, the viral enhancer can act to promote

nearby gene expression when the provirus is integrated in either orientation upstream or downstream of the host gene (Figure 1.3C; Yang et al., 2007a).

Proviral integration can also cause the expression of altered protein products. This can occur when transcription initiates from the promoter in the 5' LTR and uses the viral splice donor site to splice to exons of downstream genes (Figure 1.3D). Another mechanism by which integration can generate altered protein products is when viral transcription reads through into adjacent cellular sequences. Both of these mechanisms drive the expression of viral fusion transcripts that can potentially lead to the loss of functional or regulatory domains and thereby alter protein function (Coffin et al., 1997). In theory the integration of a provirus into the gene body can also abrogate expression of a given gene.

**Figure 1.3: Mechanisms of retroviral insertional mutagenesis.** (A) Activation of host gene expression by viral promoter insertion. The LTRs of the integrated provirus contain strong promoter elements that can drive the overexpression of nearby host genes. (B) Viral promoter can splice into downstream exons of a nearby gene to drive the expression of splice variants. (C) Viral enhancer elements contained in the LTRs can also promote nearby gene expression. (D) If the virus integrates within a gene it can lead to the expression of a truncated viral fusion transcript that may have altered function or regulation.


A
LTR
Retrovirus
LTR
Overexpression of splice variants
B
LTR
Retrovirus
LTR
Viral promoters
Endogenous promoter
Enhanced mRNA expression from 3' viral LTR promoter
C
LTR
Retrovirus
LTR
mRNA overexpression
D
LTR
Retrovirus
LTR
Expression of truncated mRNA

While there is an initial inherent integration bias, *in vivo* selective forces shape the final distribution of integration sites in the host cell genome. For instance, if a provirus integrates into or near an oncogenic or pro-survival gene, cells with this integration will become clonally expanded and form a tumor. By mapping integration sites in tumors, genes that harbor clonally expanded or recurrent integrations can be identified. This process of integration and selection allows us to find novel genes that may be involved in tumorigenesis. Previous studies have shown common integration sites in ALV-induced lymphomas in *MYC*, *MYB, BIC* (miR 155 precursor), and *TERT* genes (Hayward, Neel, and Astrin 1981; Baba and Humphries 1986; Clurman and Hayward 1989; Yang et al. 2007a; Justice, Morgan, and Beemon 2015a).

Initial studies mapping integrations in ALV-induced B-cell lymphomas found that nearly all (~80%) analyzed tumors possessed integrations upstream of *MYC* transcription start site (Hayward et al., 1981a; Neel et al., 1981). Integrations led to the production of a fusion transcript that contained a full length copy of the *MYC* RNA by a mechanism consistent with promoter insertion. Tumors containing MYC integrations were induced by infecting chickens 2-7 days post hatching with ALV. Infections produced long latency, late onset B-cell lymphomas 4-6 months post infection. Later, integrations in *BIC*, the precursor to the noncoding mir-155, were found to co-occur relatively frequently with c-myc activation and play a role in later stages of lymphomagenesis (Clurman and Hayward, 1989; Tam et al., 1997).

Further work from the same lab showed that infection of 10-day-old chicken embryos with a recombinant ALV strain, deemed EU-8, resulted in rapid onset B-cell lymphomas. Interestingly, in these tumors, the dominant integration site identified was at

the MYB locus (Kanter et al., 1988). Integrations were detected predominantly upstream of the AUG initiation codon in intron 1 and generated truncated viral fusion transcripts. The EU-8 viral strain possessed a deletion in a viral sequence that negative regulated splicing (NRS) allowing for increased efficiency of viral readthrough and viral splicing to downstream genes (Smith et al., 1997)

Our lab performed a similar insertional mutagenesis screen making use of a similar ALV variant, LR-9, with a mutation in the NRS, which induced rapid-onset B-cell lymphomas. These tumors were previously analyzed by lower-throughput methods (Polony et al., 2003; Yang et al., 2007a). The *TERT* promoter was identified as a common site of integration in these tumors with all integrations occurring between approximately 200 and 2500 bp upstream of the *TERT* transcriptional orientation. Interestingly, all identified clonally expanded integrations were in the opposite orientation of *TERT* transcription (Yang et al., 2007a). At the time, the integrations were thought to activate *TERT* expression in these tumors via insertion of the viral enhancer.

Our insertional mutagenesis screen, partially described here, improves upon the aforementioned studies by sequencing integration sites in tumors in a high throughput manner. Sequencing of a large number of tumors allows us to identify recurrent integration sites that are common to multiple independent tumors. Further, the use of random sonication to fragment the genomic DNA in the sequencing library preparation allow us to also quantify unique breakpoints (Firouzi et al., 2014; Justice et al., 2015a, 2015b). The larger number of breakpoints, the higher the abundance of a specific integration in a given tumor and thus the more clonally expanded that integration was.

Both clonal expansion and recurrent integration into a region would implicate a nearby gene in the formation of the tumor.

This study identified a number of expected targets such as *MYB* and *MYC*. The most clonally expanded integration site was the *TERT* promoter with an average of 19 breakpoints per integration, in agreement with the previous preliminary characterization of these tumors by Yang, et. al. (Justice et al., 2015b; Yang et al., 2007a) Again, consistent with previous reports from the lab, the majority of integrations were in the *TERT* promoter region in the opposite transcriptional orientation of *TERT*. In Chapter 6 of this thesis, I follow up on the role of these integrations in the *TERT* promoter in tumorigenesis.

Other common integration sites that exhibited clonal expansion were the phosphatase genes, *CTDSPL* and *CTDSPL2*. *CTDSPL* has previously been linked to cancer as a tumor suppressor gene. Relatively little was known about the function of *CTDSPL2*. In Chapter 5, I explore how integrations in these genes affects expression and how this subsequently leads to the development of cancer.

The vast majority of putative oncogenes discovered in early mutagenesis screens have proven to be important players in human tumorigenesis indicating that the chicken is a valid model for studying cancer. By identifying sites of ALV integration in B-cell lymphomas we are able to provide insights into the molecular underpinnings of initiation, progression and spread of tumors.

# Chapter 2 – Characterization of ALV integration pattern

**Summary**

Integration of the proviral genome into the host cell genome is an obligate part of the retroviral life cycle. Integration however is not a random process. Different retroviruses display distinct integration site preferences. While the integration patterns of HIV-1 and MLV have been well characterized, the pattern of ALV has only been studied with relatively few data. In this study we make use of high throughput sequencing to identify more than nine thousand ALV integration sites in the chicken genome. This is more than 10-fold the amount of data that has been analyzed in past studies. We also make use of new bioinformatics tool sets to correlate integrations with specific genomic annotations, as well as gene expression, splicing, gene length, CpG islands, and transcription start sites. We observe that ALV integration deviates slightly from what would be expected by random chance. However, we do not observe any strong integration biases such as that which is seen with MLV and HIV-1. Our data improves upon previously existing studies and sets the stage for further analysis of the mechanism by which ALV integration is targeted.

## 2.1 Introduction

Integration into the host cell genome is a defining feature of the retroviral life cycle. While it is believe that the majority of the host cell genome is available for integration, the process is not random (Engelman, 1994). There are seven known genera of retroviruses and each has unique integration site biases. For instance, lentiviruses such as human immunodeficiency virus type-1 (HIV-1) have a very strong bias for integrating into actively transcribed genes, with some reports suggesting that more than ¾ of all HIV-1 integrations are in transcribed gene units (Lewinski et al., 2006; Mitchell et al., 2004; Schröder et al., 2002). The level of transcription is also important in dictating integration events, with integration favored in more highly expressed genes (Mitchell et. al. 2004). More recently, HIV-1 was also shown to have a preference for integrating into highly spliced genes (Singh et al., 2015). The deltaretroviral genus, which includes HTLV-1, also displays a bias for integration into transcriptionally active regions (Derse et al., 2007). Murine leukemia virus (MLV), of the gammaretrovirus family, prefers to integrate near transcription start sites, CpG islands and enhancer regions (Lafave et al., 2014; Mitchell et al., 2004; De Ravin et al., 2014; Wu et al., 2003). In contrast to the aforementioned retroviral genera, integration of spumaviruses such as prototype foamy virus (PFV) is highly disfavored in or near genes or transcribed regions (Trobridge et al., 2006). The most randomly integrating retroviruses are the alpha- and betaretroviral genera which include avian leukosis virus (ALV) and mouse mammary tumor virus (MMTV) respectively (Barr et al., 2005; Faschinger et al., 2008; Mitchell et al., 2004; Narezkina et al., 2004; Withers-Ward et al., 1994).

ALV integration has been interrogated in multiple studies that have found the integration pattern to be relatively random. The first such study, Mitchell *et. al*. found that ALV integration is not significantly favored at the transcription start site, does not correlate with gene expression, and is modestly enriched near CpG islands (Mitchell et al., 2004). This study analyzed 469 integration sites in human HEK293T cells. A study from the same year, Narezkina *et. al.,* analyzed an even smaller data set, 226 integrations, also in the human genome and found that transcription units were favored integration targets and that there was a preference for actively expressed genes though extent of expression had no effect on integration frequency (Narezkina et al., 2004). Barr *et. al*. looked at integration preferences of ALV in the chicken genome and similarly found a slight preference for actively transcribed genes (Barr et al., 2005).

In each of these studies integration sites were cloned and sequenced using linker mediated inverse PCR in a low throughput manner. Thus, the maximum number of sites looked at in any one study was approximately 800. Further, gene expression data used to correlate integration with expression level was obtained using microarrays and did not include all genes. In our study, we have used high throughput sequencing to identify and map more than 9,000 integration sites, a significant improvement upon previous studies. We also have obtained RNA-seq data with expression data for all genes. Bioinformatics toolsets have also improved and allow us to correlate all integrations with various features such as transcription start sites, CpG islands, gene length, splicing and more.

We find that similar to previous reports ALV does prefer to integrate into expressed genes relative to random but that level of expression does not enhance integration. We find that there is a slight preference for integration near, but not at, the 5' end of genes. In

disagreement with some previous data we find that there is an enrichment of integration around CpG islands, though this preference is not comparable to that observed for MLV. We also see that there is a slight preference for integration flanking transcription start sites, which has not been seen before. Lastly we observe that there is a preference for integration into smaller genes as well as spliced genes.

## 2.2 Results

*ALV integration is significantly enriched into expressed genes*

Previous works looking at the integration pattern of ALV in cell culture were done with a much smaller number of integrations and a less well-annotated genome. In order to better characterize the integration pattern of ALV, we performed high throughput sequencing of integration sites in DT40 cells, a chicken B-cell lymphoma derived line (Winding et. al. 2001). Importantly, we allowed infection to proceed for only 24 hours and thus there should not be a significant amount of selection for integration sites post integration. We recovered 9,210 unique integration sites. We generated a random set of integrations by generating 350,000 random reads (RandomBed) and mapping them back to the chicken genome through the same bioinformatics pipeline used to analyze our experimental DT40 integration data. All subsequent analyses were performed on the random and experimental sets of reads in parallel. The vast majority of integrations (91%) had only one breakpoint indicating that there was very little clonal expansion of integration sites at 24 hours post infection. The lack of clonal expansion indicates a lack of selection and thus an unbiased integration profile.

After integration sites were successfully mapped to the galGal4 RefSeq genome using BowTie, they were analyzed using HOMER bioinformatics tools to assign a specific annotation to each site (Table 2.1). ALV integrations were observed to be near random for most genomic features analyzed. Our data is also consistent with previous reports of ALV integration pattern. The most notable deviation from random is the observed preference of ALV to integrate into RefSeq genes. We observed that approximately 40% of ALV integrations fell within genes compared to only 27% of

random integrations. There also appears to be a 15-fold enrichment of ALV integrations into satellite sequences as compared to random. Lastly, we observe an approximately 2-fold enrichment of ALV integration in the proximity of CpG islands. Integration pattern did not differ significantly in chick embryo fibroblasts indicating that pattern is likely not cell-type specific (S. Malhotra, data not shown).

| Annotation | Random | DT40 |
|---|---|---|
| 3UTR | 0.4 | 0.6 |
| miRNA | 0.0 | 0 |
| ncRNA | 0.0 | 0 |
| TSS (within 5kb) | 5.9 | 10.1 |
| TTS | 0.7 | 1 |
| LINE | 6.8 | 10.7 |
| SINE | 0.2 | 0.02 |
| tRNA | 0.0 | 0 |
| Exon | 0.6 | 1 |
| Intron | 12.6 | 13.8 |
| RefSeq genes | 27.0 | 40.4 |
| Promoter | 0.6 | 0.56 |
| 5UTR | 0.0 | 0.05 |
| CpG-Island | 1.1 | 2.5 |
| Low_complexity | 0.6 | 0.3 |
| LTR | 1.6 | 3.1 |
| Simple_repeat | 0.6 | 1.5 |
| Unknown | 0.0 | 0.07 |
| Satellite | 0.3 | 4.5 |

**Table 2.1: ALV integration into selected genomic annotations.** Using HOMER bioinformatics tools, we calculated the percent of integrations that fall within each of the listed genome annotations in the galGal4 reference genome. We simulated a matched random control set of integrations for comparison.

To further investigate the preference for ALV to integrate into genes, we asked whether ALV had a bias for integrating into expressed or non-expressed genes. DT40 expression levels were gathered from publically available RNA-seq data (SRR912956). Integrations were called to the nearest RefSeq gene and correlated with gene expression as measured by FPKM (fragments per kilobase per million reads; Figure 2.1). We observed a slight but highly significant depletion of integrations in or near non-expressed genes (FPKM of 0). In DT40 cells, a random integration pattern would result in approximately 17% of integrations into non-expressed genes, whereas for ALV integration we observe about 12% of integrations falling into this bin indicating that ALV has a slight bias for integrating into expressed genes, however, the level of expression did not significantly affect the extent of integration.

Because our pipeline ascribes integrations to the nearest gene, in some instances integrations can be many hundreds of kilobases away from a transcriptional unit. To address this, we repeated the above-described analysis with integrations that fall within 5 kb of a transcriptional unit. We observed a very similar distribution of integrations relative to random (data not shown).

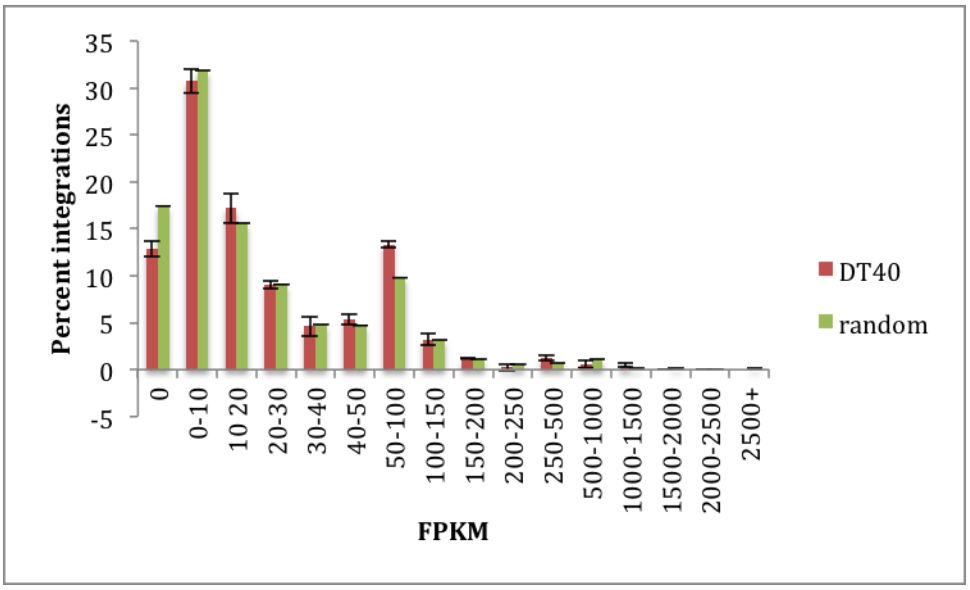
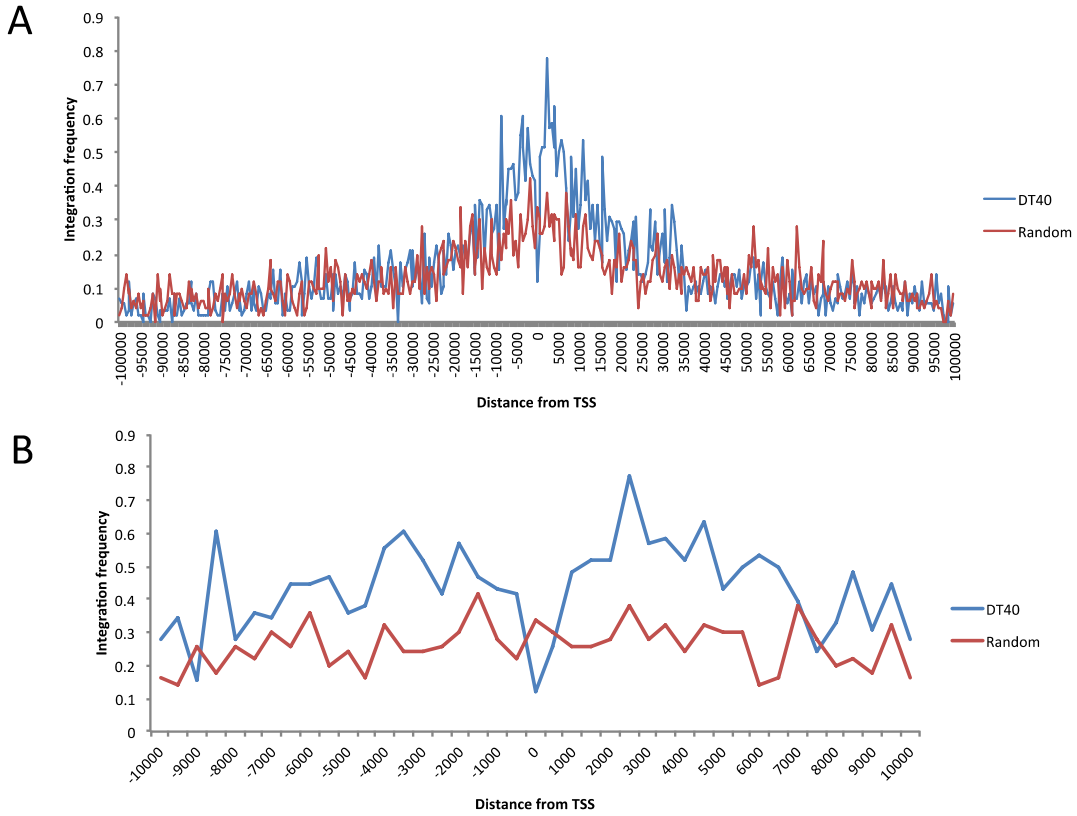



**Figure 2.1: ALV integrates preferentially into expressed genes but does not discriminate based on expression level.** We obtained DT40 transcriptome data from publically available RNA-seq data. We then generated bins of expression level and correlated integrations with the expression level of the nearest gene. Expression was quantified as FPKM (fragments per kilobase per millions reads). This analysis was also done on a matched random control set of integration sites. DT40 data is shown in red and the random control is shown in green.

*ALV has a slight preference for integration around the transcription start site of genes.*

We next looked at bias for integration in the proximity of transcription start sites (TSS). Previous ALV integration data in cell culture revealed a preference for transcribed genes, but no significant preference for the TSS specifically. In contrast, previous work from our lab has shown that there is a significant bias for integration around the TSS in ALV-induced tumor samples (Justice et. al. 2015), but it was unclear if this was impacted by selection *in vivo*. To answer this question, we again made use of HOMER bioinformatics tools to determine the distance of the site of integration to the nearest transcription start site.

We observed that ALV integration is enriched around the transcription start site, most notably in the 10 kb region flanking the TSS (Figure 2.2). Interestingly, in the immediate vicinity of the transcription start site (within 1 kb), there is a notable decrease in integration relative to random. This trend is consistent with what was previously observed by our lab *in vivo* chicken tumors, but the enrichment is much less pronounced in our cell culture system. This indicates that there is some initial bias for integrating into TSS, but these integration sites are also selected for *in vivo* making the preference appear even stronger.




**<u>Figure 2.2</u>: ALV exhibits a slight bias for integration near transcription start sites.**

Distance between the site of integration and the nearest TSS was calculated using HOMER bioinformatics tools for both DT40 as well as a matched random control. ALV integrations into DT40 cells are shown in blue, while the random control set is shown in red. Integration frequency as a function of distance from the TSS is shown in the (A) 100 kb or (B) 10 kb region flanking the TSS.

*ALV has some preference for integrating near the 5' end of genes*

Other retroviruses have been shown to exhibit preferences for integration within the gene body. For instance, MLV and HIV exhibits a strong bias for integrating into the 5' end of gene bodies (X. Wu et al. 2003, Bushman et. al. 2005). Of ALV integrations that fall into the gene body, we asked if there was a preference for position within the gene. We normalized for gene length and set up 20 equal bins throughout the gene body each corresponding to 5% of the gene length. We then determined what percent of integrations fell within each bin (Figure 2.3). Interestingly, in the first 15% of the gene body, ALV integrations into DT40s seem to be slightly enriched relative to the random control. This pattern was reproducible in CEF cells (data not shown).

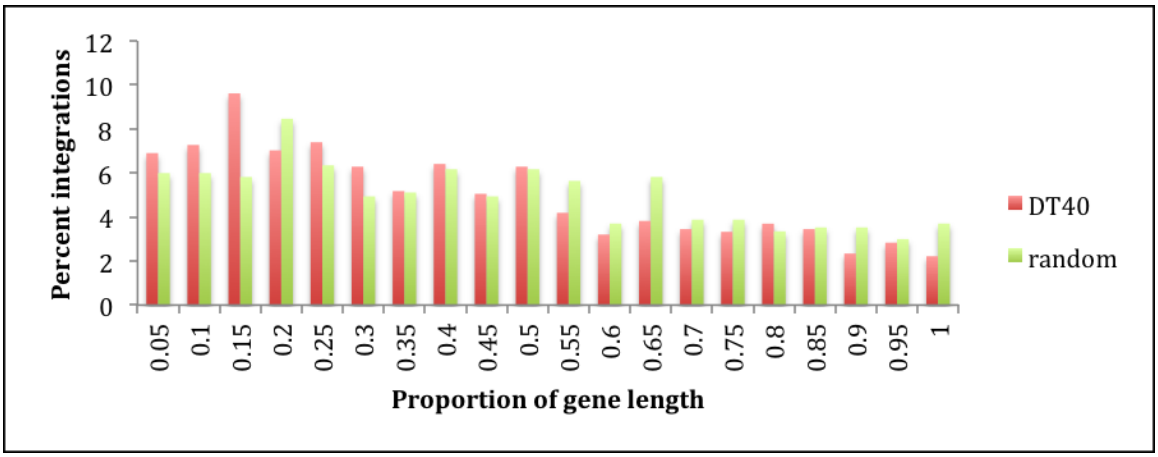



**Figure 2.3: ALV integration exhibits a slight preference for integrating near the 5' end of genes.** Integrations that fell within the body of a gene were mapped to 1 of 20 bins based on where the fell in the gene body. Shown are the percent of integrations with respect to the proportion of the gene length. Data for integrations in DT40 are shown in red and a matched random control is shown in green.

*ALV integration is biased for spliced genes.*

HIV has been shown to have a preference for integrating into highly spliced genes (Singh et al., 2015). This is a feature that has never been analyzed for ALV integration. Thus, we correlated percent of total ALV integration with the extent of splicing of the nearest gene (Figure 2.4). There was a notable depletion of integrations into unspliced genes. If integration were random, approximately 12% of the integrations would fall into unspliced genes. However, in DT40 cells, we observed that on average only 2% of integrations fell into unspliced genes. However, there did not appear to be any additional preference for spliced genes based on the extent of splicing.

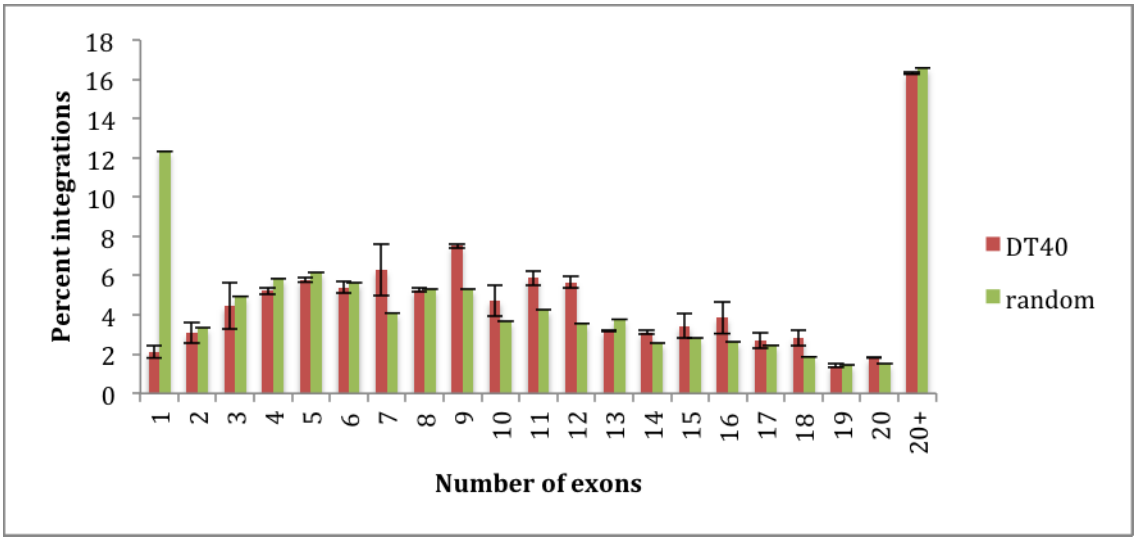



**Figure 2.4: ALV has a preference for integrating into spliced genes.** Integrations were correlated to the number of exons of the nearest RefSeq gene in the galGal4 chicken genome assembly. This analysis was performed in parallel for ALV integrations in DT40 cells as well as a matched random control set of integration sites.

*ALV prefers to integrate into shorter rather than longer genes*

We hypothesized that perhaps the preference for integrating into spliced genes could really be a preference for integrating into larger genes. Thus, we correlated integrations to the total length of the nearest gene (Figure 2.5). We observed that contrary to the preference for integrating into spliced genes, ALV actually tends to integrate into shorter rather than longer genes as compared to a matched random control.

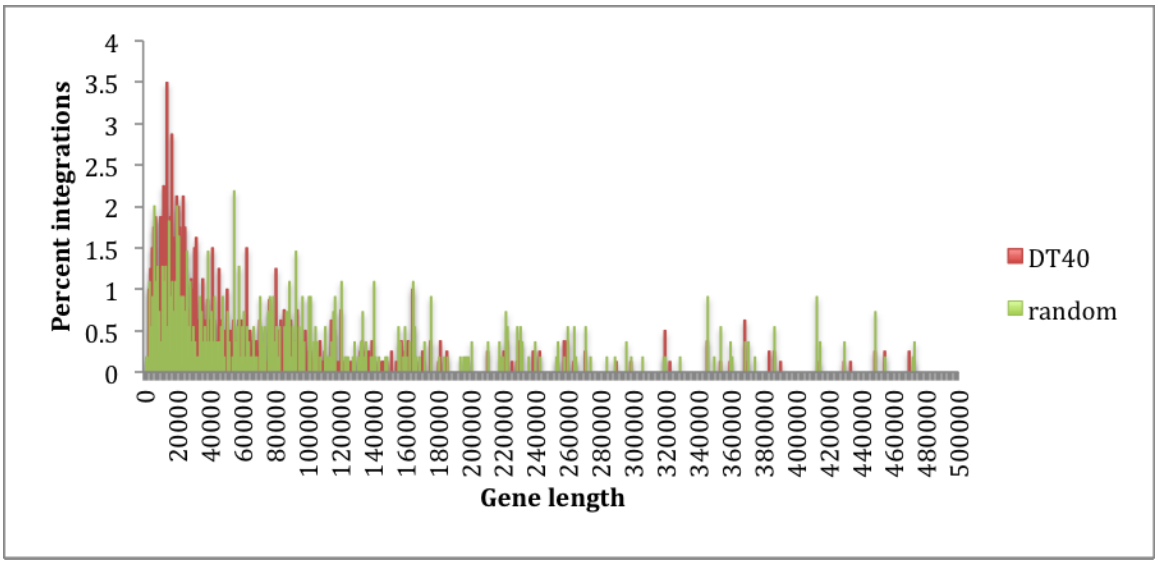



**Figure 2.5: ALV prefers to integrate into smaller genes than would be expected by random chance.** The length of the most proximal gene to the site of integration was analyzed. Shown is the percent integrations that occur with respect to total gene length. ALV integration data in DT40 cells is shown in red and a matched random control is shown in green.

*ALV exhibits a bias for integration in the proximity of CpG islands*

From the initial HOMER analysis of integration sites there was an approximate 2-fold enrichment of integrations into CpG islands as compared to a matched random control (Table 2.1). To verify this and further investigate, we calculated integrations in the proximity of CpG islands. The percent of integrations in the 1 or 5 kb region flanking CpG islands was calculated for ALV integrations in DT40s and a random set of integrations (Figure 2.6). Within 1 kb of CpG islands, we observed approximately 8% of ALV integrations as compared to 5% of integrations in the random set of integrations. The enrichment was more notable when we looked at the 5 kb region flanking CpG islands. In this region we observed approximately 20% of random integrations as opposed to more than 35% of ALV integrations in DT40 cells. Preference for integration into CpG islands is strongly associated with MLV integration (Wu et. al. 2003) but we report here that ALV also has a fairly strong preference for integrating in the vicinity of CpG islands though the enrichment is not as high as that observed for MLV.

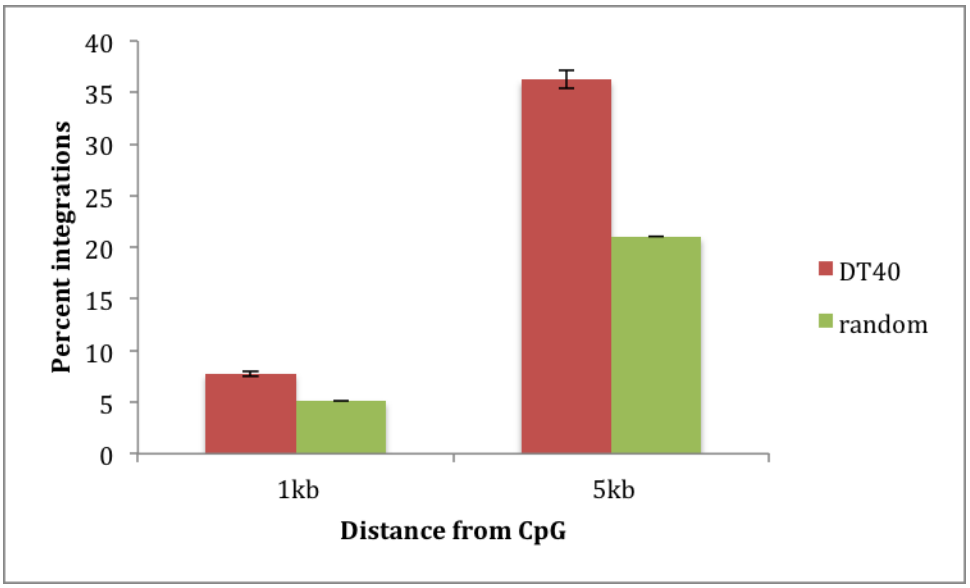



**Figure 2.6: ALV has a preference for integrating near CpG islands.** CpG island locations were extracted from the galGal4 reference genome. BedWindow was used to calculate the overlap of the integration sites with CpG islands with a 1 or 5 kb range flanking each CpG island. Shown is the percent integrations that fall in the 1 or 5 kb surrounding CpG islands. Percent of ALV integrations in DT40 cells in each category are shown in red, the matched random control is shown in green.

*There is a slight sequence preference at the site of ALV integration.*

As has been previously observed in human and chicken cells, there is a slight sequence preference at the site of ALV integration. There is a strong preference for a T -3 nucleotides from the integration site as well as a preference for G at +1 and A at +9. The consensus sequence is palindromic and appears to be more symmetrical than has been previously reported *in vivo* (Figure 2.7; Justice et al., 2015b; Kirk et al., 2016).

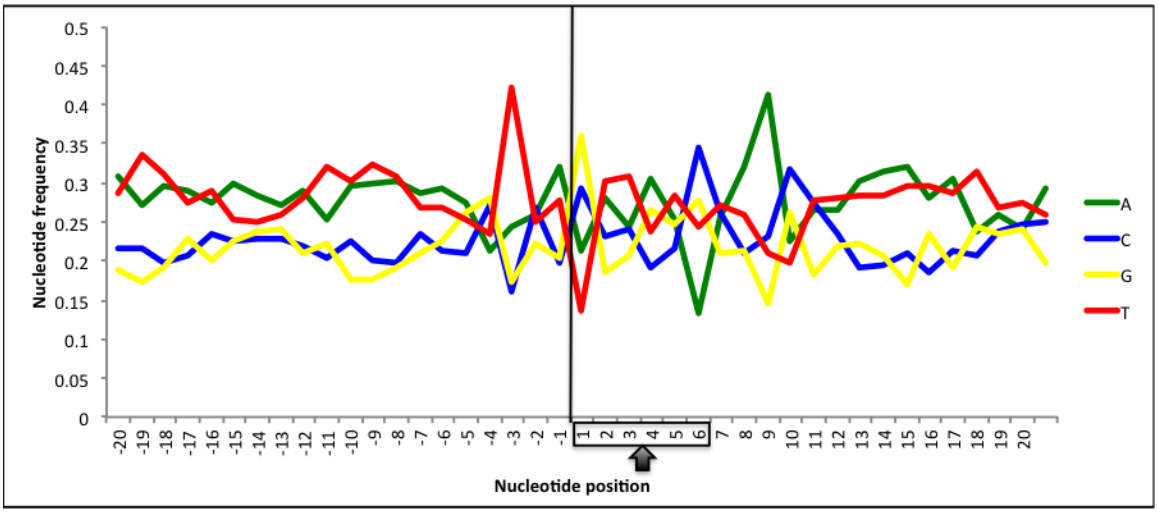



**Figure 2.7: ALV has a modest sequence preference at the site of integration.** Nucleotide frequencies flanking the site of integration were calculated using HOMER bioinformatics tools. Shown is the frequency with which each nucleotide appears at a given position flanking the integration site. The black line indicates the site of integration. The boxed nucleotide positions correspond to the sequence duplicated in the process of integration. The black arrow shows the axis of symmetry for the palindromic motif.

## 2.3 Discussion

The data reported here are largely consistent with previously published data. For instance, we find a preference for integration into genes relative to random in agreement with Narezkina *et. al*. and Barr *et. al*. . We also find that there is a preference for transcribed genes but that level of transcription does not significantly affect extent of integration consistent with Narezkina *et. al*. However, unlike previous studies, we observed a preference for integration around the transcription start site of genes as well as in the proximity of CpG islands. In addition we correlated integration with gene splicing, gene length, and gene body, all attributes that had not been previously characterized. We see a preference for integration into short genes, spliced genes and the 5' end of genes. Our data may differ from previous studies largely due to sample size, a better-annotated genome and better bioinformatics tools.

While DT40 cells serve as the parental "wild-type" condition in our studies in the following chapters, they were actually originally isolated from a bursal lymphoma. Thus, it was unclear whether the integration pattern observed in DT40 cells would be more similar to that observed in primary cell lines or tumors. A parallel study in the lab analyzed integration pattern of ALV in primary chicken embryo fibroblasts in culture as well as in chicken tumors (Malhotra et al., 2017). It was found that the largest discrepancy between integration pattern in primary cells and tumors was in the proximity of transcription start sites and CpG islands. In tumors, integration was more highly preferred near the transcription start site and less preferred in the proximity of CpG islands as compared to the integration pattern in primary cells. The integration pattern described here for DT40 cells agrees very closely with that observed in primary chick

embryo fibroblasts. This indicates that the discrepancies in integration pattern seen in tumors are the result of selection over time. Thus, DT40 cells are a relevant model in which to study wild type ALV integration pattern.

While we do see slight integration biases, ALV integration is still significantly more random than either MLV or HIV-1 integration. For instance, we see approximately 40% of ALV integration into genes as compared to 27% for a matched random control. While this is certainly enriched above random, it is not as strong of a preference as HIV-1. Nearly 75% of HIV-1 integration have been reported to fall within transcribed genes (Mitchell et al., 2004; Schröder et al., 2002). Similarly, depending on the study, an average of about 12% of MLV integrations fall within 1 kb of CpG islands, compared to 5% expected by random chance. We observe only roughly 7% of ALV integration in this same region, an increase above random, but not a strong bias as observed for MLV (Mitchell et al., 2004). We see about 10% of total ALV integrations fall within 5 kb of a transcription start site compared to only about 5% of random integrations. However, more than 25% of MLV integrations fall in that same category (Wu et al., 2003). Thus, while ALV does deviate from a random integration pattern, there are no striking integration biases.

It was originally thought that integration might be influenced by chromatin accessibility. However, the distinct integration site preferences of different retroviruses make it unlikely that this is the sole mechanism determining integration site selection. Early studies on yeast Ty retrotransposons, which are similar to retroviruses, discovered that integration is targeted to very specific chromosomal locations by tethering of the integration complex to the chromatin by host cell factors (Lesage and Todeschini, 2005).

For instance, binding of the integrase protein to the Sir4 protein targets the Ty5 transposon to heterochromatin (Dai et al., 2007). This led to screens looking for similar host cell factors that might be responsible for retroviral targeting. Such studies have identified LEDGF and CPSF6 as targeting factors of HIV-1 and the BET proteins as targeting factors for MLV integration (Maertens et al., 2003; Sharma et al., 2013; Sowd et al., 2016).

The majority of human gene therapy trials to date have utilized HIV-1- and MLV-based vectors, which due to their biased integration into genes and regulatory regions have a high incidence of insertional mutagenesis (Hacein-Bey-Abina et al., 2008; Howe et al., 2008). The more random integration pattern of ALV integration may reduce the risk of deleterious integration (Suerth et al., 2014). Understanding how ALV integration is regulated could facilitate the development of ALV-based vectors for use in human gene therapy.

# Chapter 3 – Characterization of host cell factors that regulate ALV integration: FACT complex

Adapted from:
**Winans, S**., Larue, R.C., Abraham, C.M., Shkriabai, N., Skopp, A., Winkler. D., Kvaratskhelia, M., and Beemon, K.L. (2017) The FACT complex promotes avian leukosis virus DNA integration. *Journal of Virology* 91(7). pii: e00082-17.

**Summary**

All retroviruses need to integrate a DNA copy of their genome into the host chromatin. Cellular proteins regulating and targeting lentiviral and gammaretroviral integration in infected cells have been discovered, but the factors that mediate alpharetroviral avian leukosis virus (ALV) integration are unknown. Here, we have identified the FACT protein complex, which consists of SSRP1 and Spt16, as a principal cellular binding partner of ALV integrase. Biochemical experiments with purified recombinant proteins show that SSRP1 and Spt16 are able to individually bind ALV IN, but only the FACT complex effectively stimulates ALV integration activity *in vitro*. Likewise, in infected cells, the FACT complex promotes ALV integration activity with proviral integration frequency varying directly with cellular expression levels of the FACT complex. An increase in 2-LTR circles in the depleted FACT complex cell line indicates that this complex regulates the ALV life cycle at the level of integration. This regulation is shown to be specific to ALV, as disruption of the FACT complex did not inhibit either lentiviral or gammaretroviral integration in infected cells. Integration pattern was subtly but significantly affected by knockdown of the FACT complex indicating that the complex may also play a role in integration site selection.

## 3.1 Introduction

Retroviral integration into the host genome is not random and varies dramatically across genera. Lentiviral HIV-1 has been shown to exhibit strong integration site preferences within active gene units, whereas gammaretroviral MLV exhibits a strong preference for enhancers and transcription start sites (Lewinski et al., 2006; Schröder et al., 2002; Wu et al., 2003). These biases have been attributed to interaction of IN in the context of the pre-integration complex with their cognate host cell factors (Craigie and Bushman, 2014; Debyser et al., 2015; Kvaratskhelia et al., 2014). For example, cellular chromatin associated protein lens epithelial-derived growth factor (LEDGF/p75) interacts with HIV-1 IN and directs lentiviral integration into actively transcribed genes (Ciuffi et al., 2005; Llano et al., 2006; Maertens et al., 2003). Similarly, BET (bromodomain and extraterminal domain) proteins have been shown to interact with MLV IN and target MLV integration to transcription start sites, enhancers and gene regulatory regions (Aiyer et al., 2014; El Ashkar et al., 2014; Gupta et al., 2013; De Rijck et al., 2013; Sharma et al., 2013). These host cell factors bind their cognate viral IN and select histone marks to act as a bimodal tether to recruit the pre-integration complex to specific genomic regions surrounding the host factor binding sites (Craigie and Bushman, 2014; Debyser et al., 2015; Eidahl et al., 2013; Kvaratskhelia et al., 2014; Larue et al., 2014). In addition to targeting integration events to specific genomic features, these factors also serve to significantly enhance integration efficiencies (Llano et al., 2006; Sharma et al., 2013; Shun et al., 2007).

A recent study has identified cellular serine/threonine protein phosphatase 2A (PP2A) as a selective binding partner of deltaretroviral (human T cell lymphotropic virus

type 1 and 2 and bovine leukemia virus) INs (Maertens, 2016). Furthermore, the B' subunit of PP2A has been shown to bind and stimulate concerted integration of deltaretroviral INs *in vitro* (Maertens, 2016). However, unlike LEDGF/p75 and BET proteins, PP2A does not directly engage chromatin, and it remains to be seen whether this cellular protein can modulate delta-retroviral integration in infected cells.

Alpharetroviruses such as ALV exhibit a distinct integration pattern with seemingly random distribution of integration sites throughout chromatin and with only a slight preference for integrating into gene regions (Barr et al., 2005; Mitchell et al., 2004; Narezkina et al., 2004; Withers-Ward et al., 1994). To understand how ALV integration is regulated by host cellular factors, we have performed affinity capture of the ALV IN protein followed by mass spectrometry (MS)-based proteomics experiments to identify protein binding partners. Using this approach we identified structure specific recognition protein 1 (SSRP1) and suppressor of Ty 16 (Spt16), the components of the heterodimeric FACT (facilitates chromatin transcription) complex (Orphanides et al., 1999), as the top protein hits that specifically bound to ALV but not HIV-1 IN.

The FACT complex is a highly conserved general histone chaperone protein that is essential for transcription and DNA replication (Abe et al., 2011; Belotserkovskaya and Reinberg, 2004; Orphanides et al., 1998). The complex has also been shown to play important roles in DNA damage responses, centromere deposition, recombination and DNA methylation (Ikeda et al., 2011; Kumari et al., 2009; Okada et al., 2009; Oliveira et al., 2014). The FACT complex is thought to destabilize the histone octamer, providing access to the DNA for various enzymes (Formosa; Reinberg and Sims, 2006; Winkler

and Luger, 2011). The complex is also important for reassembling nucleosomes after polymerases have moved through the DNA to establish new chromatin (Formosa 2012).

In this report, we show that both components of the FACT complex, SSRP1 and Spt16, can individually bind ALV IN. Furthermore, we show that the C-terminal domain (CTD) of ALV IN is essential for the interaction with the FACT complex. *In vitro* integration activity assays reveal that the FACT complex, rather than its individual components, specifically stimulates ALV but not HIV-1 IN activity. Our findings also indicate that the FACT complex regulates ALV integration in infected cells as the frequency of ALV proviral integration is directly correlated with the abundance of the FACT complex. The decrease in proviral integration when the FACT complex was depleted was accompanied by an increase in 2-LTR circles, indicating that the FACT complex stimulates the integration step of the viral life cycle. Moreover, we show that the FACT complex specifically promotes ALV integration, as cells with depletion of the FACT complex had no inhibitory effect on either HIV-1 or MLV integration efficiencies. High throughput sequencing of viral integration sites in the presence and absence of the FACT complex reveal subtle but significant differences in pattern of integration suggesting that the FACT complex may also be playing a role in targeting ALV integration to specific genomic locations.

## 3.2 Results

*The FACT complex specifically interacts with and stimulates catalytic activity of ALV IN*

To identify host cell factors that bind ALV IN, we performed affinity capture coupled with MS analysis using recombinant His-tagged ALV and HIV-1 INs as baits and nuclear extracts of uninfected chicken DT40 and human Sup-T1 cells. Unique hits that were reproducible in both cell lines were identified through semiquantitative analysis of peptide spectral counts. This revealed SSRP1 and Spt16, the components of the FACT complex, to be main binding partners of ALV IN (Figure 3.1A). Using a taxonomy of "Homo sapiens" (human) allowed us to identify these proteins from Sup-T1 cells. However, since the MASCOT search engine does not contain a chicken taxonomy, we used higher order classification of "bony vertebrates" for analyzing DT40 proteins. In these samples Spt16 (which is not annotated in chicken cells) was identified due to the high homology with its human counterpart. The confidence in the correct identification of Spt16 in DT40 cells is high due to identification of 16 unique peptides and 18% coverage of the protein (Figure 3.1B). SSRP1 is functionally annotated in both species and was identified by MASCOT as chicken or human origin depending on the cell type. No interacting peptides from either of these FACT complex proteins were detected in parallel HIV-1 IN pull-down fractions. In contrast, as expected LEDGF/p75 peptides were detected in HIV-1 but not with ALV IN pull-downs (Figure 3.1A).

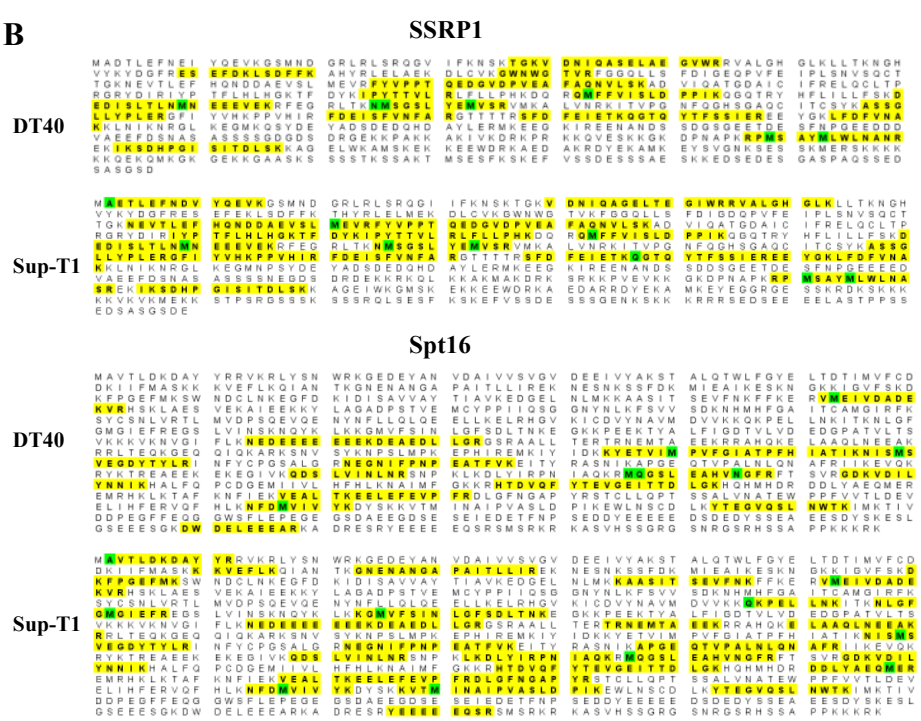
Other candidate factors that were enriched in the ALV IN pulldown fraction relative to HIV-1 are shown in Table 3.1.

**Figure 3.1: MS-based proteomics analysis of cellular binding partners of ALV and HIV-1 INs**. Two independent experiments with Sup-T1 (human) and DT40 (chicken) cells were performed. (A) Shown is the list of top unique protein hits (compiled from both cell lines) from nuclear extracts of DT40 or Sup-T1 cells. Semiquantitative values of peptide spectral counts for each identified protein are indicated. ND: no peptides from the indicated protein were detected. (B) Summary of identified peptides sequences from SSRP1 and Spt16, which are indicated in bold and highlighted in yellow. Total spectral counts for SSRP1 peptides were 85 and 32 in SupT1 and DT40 cells respectively, yielding 36% (254 out of 709) and 19% (192 out of 1006) amino acid coverage. In sharp contrast, no SSRP1 peptides were detected in parallel experiments with HIV-1 IN. Total spectral counts for Spt16 peptides in SupT1 and DT40 cells were 103 and 39, yielding 33% and 18% (341 out of 1047) amino acid coverage, respectively. Spt16 in DT40 cells was identified based on its homology with the corresponding human gene. In sharp contrast no Spt16 peptides were detected in parallel experiments with HIV-1 IN. Oxidation or other modifications are highlighted in green.

**A**

| | Cells | | | |
|---|---|---|---|---|
| | DT40 | | Sup-T1 | |
| Candidate proteins | ALV IN | HIV-1 IN | ALV IN | HIV-1 IN |
| Fact complex subunit Spt16 | 39 | ND | 103 | ND |
| Fact complex subunit SSRP1 | 32 | ND | 85 | ND |
| LEDGF/p75 | ND | 103 | ND | 15 |

**B**



| Protein name | Peptide hits | |
|---|---|---|
| | ALV IN | HIV-1 IN |
| SPT16 | 103 | <1 |
| SSRP1 | 85 | <1 |
| BRD2 | 41 | <1 |
| UBTF | 114 | 3 |
| SRRM2 | 33 | <1 |
| NCL, Nucleolin | 62 | 8 |
| SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 5 | 21 | 3 |
| HNRNPC | 33 | 12 |
| SMARCC1 | 37 | 18 |
| CDC2L1 | 65 | 35 |
| TOP1 | 50 | 28 |
| SMARCA4 | 46 | 25 |
| CHD1 | 83 | 53 |

**Table 3.1: Host cell factors that bind ALV integrase protein**. Top host cell factors that were enriched in the ALV IN pulldown fraction versus the HIV-1 IN pulldown fraction are shown. The list is sorted by the ratio of peptides detected in ALV IN pulldown vs. HIV-1 IN pulldown.

To validate our MS-based results, we next analyzed the affinity pull-down fractions by immunoblotting using antibodies directed against SSRP1 or Spt16 proteins. The results in Figure 3.2A show that ALV IN interacted with both components of the endogenous FACT complex from nuclear extracts of HEK293T cells. In contrast, in parallel reactions HIV-1 and MLV INs failed to interact with either SSRP1 or Spt16 (Figure 3.2A). Figure 3.2B shows the recombinant purified IN proteins used in Figure 3.2A.

As our MS-based results and immunoblotting cannot distinguish between direct and indirect interactions; we next performed affinity pull-downs with recombinant purified GST-tagged ALV and HIV-1 IN proteins. For these experiments, we used either purified recombinant FACT complex or LEDGF/p75. The FACT complex specifically interacted with ALV IN but not with HIV-1 IN (Figure 3.2C). In parallel experiments the expected interaction of HIV-1 IN with its known cellular cofactor, LEDGF/p75, but not the FACT complex was seen (Figure 3.2C).

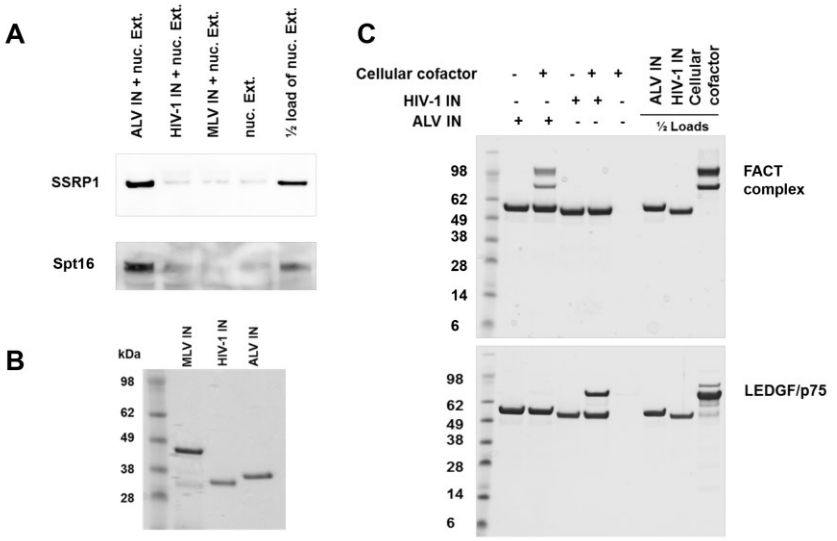
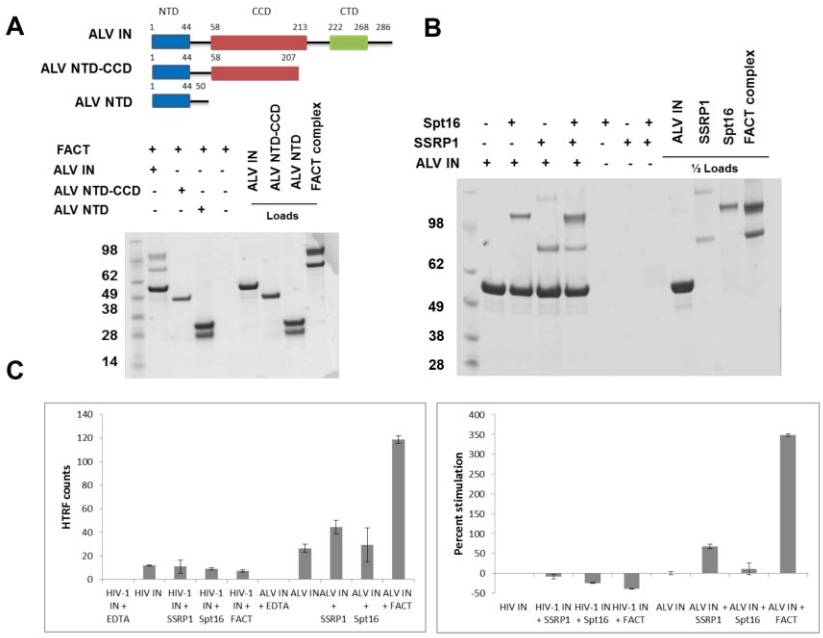



**Figure 3.2: The components of the FACT complex, SSRP1 and Spt16, bind ALV IN but not HIV-1 or MLV INs.** (A) Affinity pull-down results from uninfected nuclear lysates of HEK293T cells (100 µg total protein) with indicated 2 µM 6xHis-tagged recombinant retroviral INs, followed by immunoblotting with SSRP1 or Spt16 antibodies. (B) Coomassie-stained SDS/PAGE gel of recombinant purified INs used in panel A. (C) Coomassie-stained SDS/PAGE gel of affinity pull-down results of recombinant purified GST-tagged INs (1 µM) with FACT complex (0.6 µM). All images show representative results of triplicate experiments with molecular weights indicated.

We next wanted to further dissect the contributions of individual proteins and/or domains responsible for interaction between the FACT complex and ALV IN. We first examined binding of C-terminally truncated fragments of ALV IN with the FACT complex. The results in Figure 3.3A show that the CTD (consisting of amino acids 208-286) is essential for binding to the FACT complex as the isolated N-terminal domain (NTD) and the two domain fragment containing NTD and catalytic core domain (CCD) fail to bind the FACT complex.

To elucidate contributions of individual components of the FACT complex for binding and catalytic activity of ALV IN, we next utilized affinity pull-downs and homogenous time resolved fluorescence (HTRF)-based *in vitro* integration assays (Sharma et al., 2013). Figure 3.3B shows that both purified proteins are able to bind ALV IN individually. However, both components of the FACT complex were needed to effectively stimulate ALV integration activity (~350 percent). In contrast, SSRP1 or Spt16 alone failed to enhance ALV IN activity (Figure 3.3C). The level of stimulation of ALV IN activity by the FACT complex is similar to that seen for the addition of LEDGF/p75 to HIV-1 IN or Brd4 to MLV IN (Sharma et al., 2013). In parallel experiments, HIV-1 IN activity was not significantly stimulated by the addition of SSRP1, Spt16 or the FACT complex. In fact, the addition of either protein suppressed HIV-1 IN activity (Figure 3.3C).




**<u>Figure 3.3</u>: FACT complex stimulates in vitro integration activity of ALV integrase.** (A) Schematic of C-terminally truncated constructs of ALV IN with domains and flexible linkers indicated. Coomassie-stained SDS/PAGE analysis of affinity pull-down fractions using recombinant purified GST-tagged ALV IN, ALV NTD-CCD, and ALV NTD (1 μM) with FACT complex (0.6 μM). Lower band in ALV NTD preparation was GST alone. (B) Coomassie-stained SDS/PAGE analysis of affinity pull-down fractions using recombinant purified GST-tagged ALV IN (1 μM) with FACT complex, SSRP1 or Spt16 (0.6 μM). All images depict representative results of triplicate experiments with molecular weights indicated. (C) HTRF strand transfer integration activity assay of HIV-1 or ALV INs (400 nM) with FACT complex, SSRP1 or Spt16 (1.0 μM). The results from triplicate experiments with standard deviations are indicated. Shown are the HTRF raw counts and the percent stimulation.

*ALV proviral integration frequency correlates directly with FACT complex expression levels in infected cells.*

Since the FACT complex binds ALV IN and stimulates its activity *in vitro* (see previous), we hypothesized that the FACT complex could also play a role in regulating ALV integration in infected cells. To determine the effect of the FACT complex on ALV integration in infected cells, we employed a chicken cell line (DT40) with varying expression levels of the FACT complex. Previous research has shown that the expression and abundance of the FACT complex is regulated by a complex feedback loop in which *SSRP1* mRNA plays a critical role (Safina et al., 2013). The presence of *SSRP1* mRNA is essential for stability of Spt16 protein and the FACT complex as a whole. In the absence of *SSRP1* mRNA, both protein components are depleted. Similarly, when *SSRP1* mRNA is overexpressed, Spt16 protein levels also increase (Safina et al., 2013).

We used a SSRP1 conditional knock-out engineered in the chicken B-cell line, DT40, to investigate the functional consequences of the FACT complex on ALV integration. This cell line lacks both endogenous copies of the *SSRP1* gene but has a wild type *SSRP1* gene expressed from a tet-repressible promoter (*SSRP1*$^{-/-}$ + *SSRP1*) (Abe et al., 2011) (see also Figure 3.4A). Because of the demonstrated complex regulation of the FACT complex, this cell line, which allowed us to manipulate *SSRP1* levels, is ideal for controlling the levels of SSRP1, Spt16 and the FACT complex as a whole.

In the presence of doxycycline, (*SSRP1*$^{-/-}$) cells exhibited SSRP1 protein levels that declined to undetectable levels by 12 hours post treatment, resulting in a cell line with no functional FACT complex (Figure 3.4B). Of note, FACT complex knockdown

did not significantly affect cell proliferation during the initial 48 hours after doxycycline addition (Figure 3.4C).

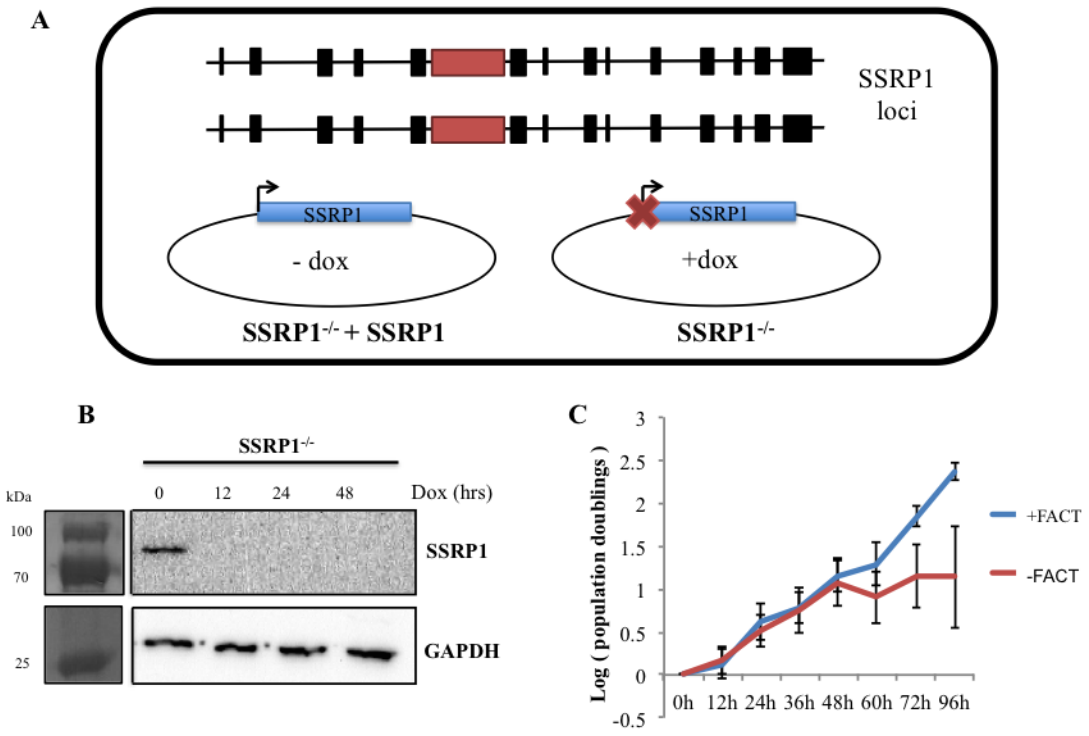



**Figure 3.4: Validating SSRP1 conditional knockout cell line**. (A) Schematic of cell line. In a wild type DT40 background, both endogenous loci of *SSRP1* were targeted by homology constructs (indicated by red box) in order to knock down both copies of the gene. A wild type copy of the *SSRP1* gene was introduced into cells on a plasmid under the control of a tet-repressible promoter. In the absence of doxycycline, *SSRP1* is expressed (*SSRP1*$^{-/-}$ + *SSRP1*). In the presence of doxycycline, *SSRP1* expression is ablated (*SSRP1*$^{-/-}$). (B) Western blot showing that SSRP1 protein levels decrease to undetectable levels by 12 hours after doxycycline addition. Relevant molecular weight markers are shown. (C) Growth curve of SSRP1 knockout and wild type cells over 96 hour period after adding doxycycline. *SSRP1* knockout (*SSRP1*$^{-/-}$) does not significantly affect cell growth and proliferation until after 48 hours.

To analyze how varying levels of SSRP1 could affect ALV integration, we compared infections in parental DT40 cells, cells expressing elevated levels of *SSRP1* (*SSRP1*$^{-/-}$ + *SSRP1*) or knockout (*SSRP1*$^{-/-}$) levels of *SSRP1*. In the manipulated cell line (*SSRP1*$^{-/-}$ + *SSRP1*), *SSRP1* is expressed from an exogenous promoter and is thus not expressed at wild type levels. We observed a 5-fold increase in *SSRP1* mRNA expression in *SSRP1*$^{-/-}$ + *SSRP1* relative to the parental DT40 cell line.

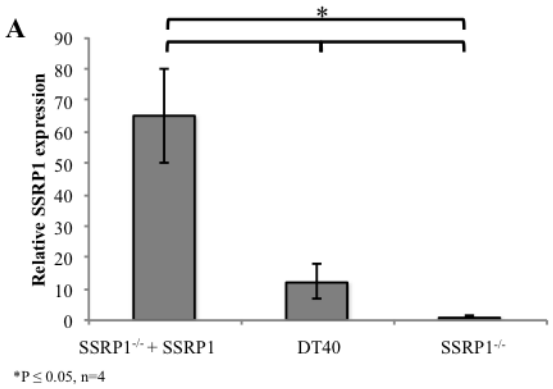
*SSRP1*$^{-/-}$ cells express 65- and 10-fold lower *SSRP1* mRNA compared with *SSRP1*$^{-/-}$ + *SSRP1* and parental DT40 cells respectively (Figure 3.5A). By using these three conditions, we can assay for ALV proviral integration frequency at wild type levels, overexpressed and knockout levels of *SSRP1* and hence the levels of the FACT complex.

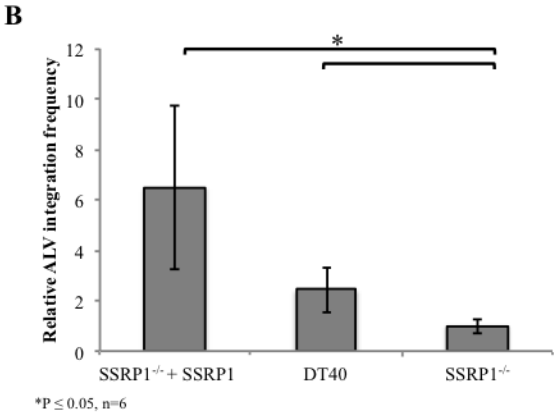
Cells expressing the highest levels of SSRP1 also had the highest levels of ALV integration frequency, determined by qPCR analysis of gel-purified DNA**.** We observed an approximately 2-fold increase in proviral integrations in the wild type DT40 cells versus knockout cells (*SSRP1*$^{-/-}$) and a 6-fold increase in integration frequency in cells overexpressing SSRP1 (*SSRP1*$^{-/-}$ + *SSRP1*) (Fig 3.5B*)*. Thus, the trend in integration frequency directly correlates with expression levels of *SSRP1* as well as the levels of the FACT complex.

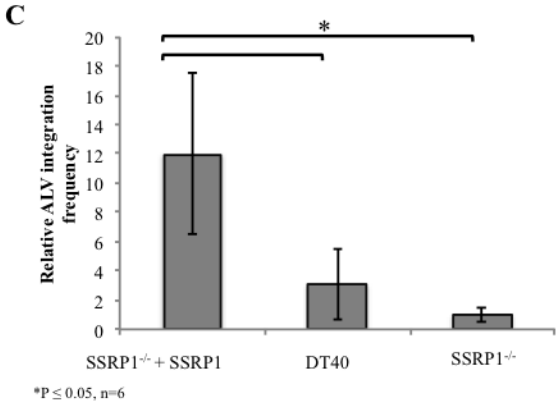
These trends were verified using a second, independent method. In this approach, proviral integration frequency was measured from genomic DNA collected from infected cells using nested PCR. A first round of PCR was performed to enrich for proviral-host genome junctions using a viral specific primer and a consensus primer within the most abundant repeat element in the chicken genome (CR1 element). A second, quantitative PCR (qPCR) was then performed using primers within the virus. This method confirmed

a significant 12-fold decrease in proviral integrations in infected *SSRP1* knockout cells relative to *SSRP1*$^{-/-}$ + *SSRP1* cells (Figure 3.5C). The parental DT40 cell line had an intermediate level of integration consistent with *SSRP1* expression levels. These data show that ALV proviral integration frequency is directly correlated to FACT complex abundance and indicate that the FACT complex promotes ALV integration in infected cells.

**Figure 3.5: ALV proviral integration frequency correlates directly with SSRP1 mRNA expression levels.** (A) SSRP1 expression in parental DT40 cells, *SSRP1*$^{-/-}$, and *SSRP1*$^{-/-}$ + *SSRP1* cells. The expression levels of *SSRP1* mRNA were measured relative to a housekeeping gene, *RPL30* by qRT-PCR. In the knockout condition (*SSRP1*$^{-/-}$), *SSRP1* expression decreased 65- and 10-fold relative to *SSRP1*$^{-/-}$ + *SSRP1* and parental DT40 cells respectively (n=6, $p < 0.05$). DT40 expression was approximately 6-fold lower than expression levels in *SSRP1*$^{-/-}$ + *SSRP1* cells (n=6, $p < 0.05$). (B) Analysis of integration frequency. Proviral integrations were measured by qPCR from gel purified genomic DNA. In the *SSRP1* knockout cell line, proviral integration frequency decreased 6.5-fold relative to cells expressing *SSRP1* (*SSRP1*$^{-/-}$ + *SSRP1*) (n=6, $p<0.05$). DT40 cells exhibited approximately 2.5 fold higher expression than knockout cells (n=6, $p<0.05$). (C) Proviral integrations were measured independently using a CR1-gag nested PCR approach. Integration frequency decreased by 12-fold in the absence of *SSRP1* expression (*SSRP1*$^{-/-}$) relative to *SSRP1*$^{-/-}$ + *SSRP1* cells (n=5, $p<0.005$).


A
*
Relative SSRP1 expression
90
80
70
60
50
40
30
20
10
0
SSRP1-/- + SSRP1
DT40
SSRP1-/-
*P ≤ 0.05, n=4


B
*
Relative ALV integration frequency
12
10
8
6
4
2
0
SSRP1-/- + SSRP1
DT40
SSRP1-/-
*P ≤ 0.05, n=6


C
*
Relative ALV integration frequency
20
18
16
14
12
10
8
6
4
2
0
SSRP1-/- + SSRP1
DT40
SSRP1-/-
*P ≤ 0.05, n=6

*The FACT complex regulates ALV at the level of integration*

Because the proteins of the FACT complex interacted with ALV IN, we hypothesized that this complex is specifically regulating the ALV life cycle at the level of integration. However, the observed change in the number of detectable proviral integration events could be due to effects of the FACT complex at various levels of the retroviral life cycle preceding or during integration. For instance, if the FACT complex affects reverse transcription, nuclear import or integration, then one would expect to detect less integrants in the *SSRP1* knockout (*SSRP1*$^{-/-}$) cells.

To distinguish between these possibilities, we quantified various retroviral intermediates. Plus strand extension (PSE) products are an intermediate of the retroviral life cycle produced by the late steps of reverse transcription and the abundance of PSE products can be used to assay variations in reverse transcription (Karageorgos et al., 1995). There was no significant difference in the levels of PSE products between the knockout (*SSRP1*$^{-/-}$) cells versus cells expressing *SSRP1* (*SSRP1*$^{-/-}$ + *SSRP1*) indicating that reverse transcription is not affected by the levels of the FACT complex (Figure 3.6A).

Once viral cDNA is reverse transcribed, it enters the nucleus of the host cell as part of the pre-integration complex. Within the nucleus, the non-homologous end-joining pathway circularizes unintegrated viral genomic DNA to generate 2-LTR circles. These circularized viral intermediates can be used as a proxy to measure the abundance of unintegrated nuclear viral genomes (Butler et al., 2001; Mandal and Prasad, 2009). The unique LTR-LTR junction present in this intermediate makes it readily detectable and distinguishable by PCR (Butler et al., 2001). A 10-fold increase in 2-LTR circles was

detected in the *SSRP1* knockout (*SSRP1*$^{-/-}$) cells versus cells expressing *SSRP1* (*SSRP1*$^{-/-}$ + *SSRP1*) (Figure 3.6B). A decrease in proviral integration accompanied by an increase in unintegrated nuclear viral products, specifically 2-LTR circles, indicates that nuclear import is not blocked and that the integration step is significantly impaired in the absence of the FACT complex.

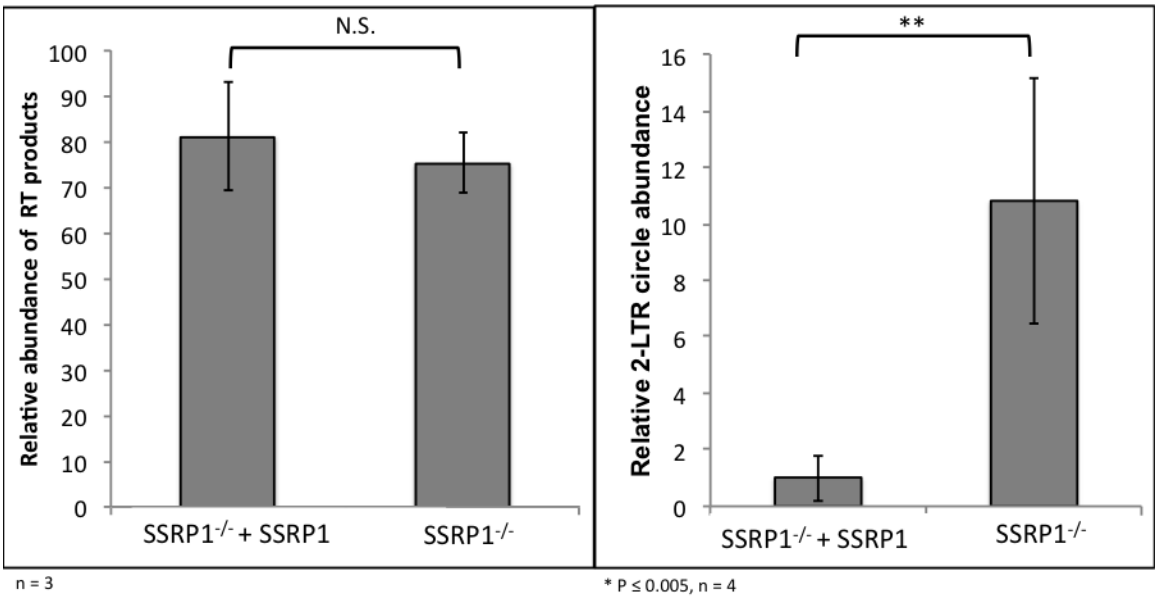



**Figure 3.6: The FACT complex promotes ALV integration.** To determine the step of the life cycle that the FACT complex is affecting, retroviral intermediates were assayed by qPCR. (A) The FACT complex does not disrupt reverse transcription. Abundance of plus strand extension (PSE) products, a product of late reverse transcription, was measured by qPCR using primers within *gag*. There was no significant difference in PSE product abundance observed between *SSRP1* knockout (*SSRP1*$^{-/-}$) and cells expressing *SSRP1* (*SSRP1*$^{-/-}$ + *SSRP1*) (n=3). (B) The FACT complex specifically promotes ALV integration. The abundance of 2-LTR circles was measured by qPCR in cells expressing *SSRP1* (*SSRP1*$^{-/-}$ + *SSRP1*) and *SSRP1* knockout (*SSRP1*$^{-/-}$) cells. A 10-fold increase in 2-LTR circles was detected in the *SSRP1* knockout cells (n=4, $p<0.005$) indicating that depletion of the FACT complex inhibits integration.

*Knockdown of the FACT complex does not inhibit lentiviral or gamma-retroviral integration*

We next wanted to know if the regulation of integration by the FACT complex was specific to ALV or could also affect other retroviruses. As such, we infected knockout cells (*SSRP1*$^{-/-}$) or cells overexpressing *SSRP1* (*SSRP1*$^{-/-}$ + *SSRP1*) with VSV-G pseudotyped MLV or HIV-1. There was no significant difference in the frequency of MLV or HIV-1 proviral integration in the knockout (*SSRP1*$^{-/-}$) cells relative to cells expressing *SSRP1* (*SSRP1*$^{-/-}$ + *SSRP1*) (Figure 3.7). The levels of plus strand extension products did not significantly differ nor did abundance of 2-LTR circles (data not shown).

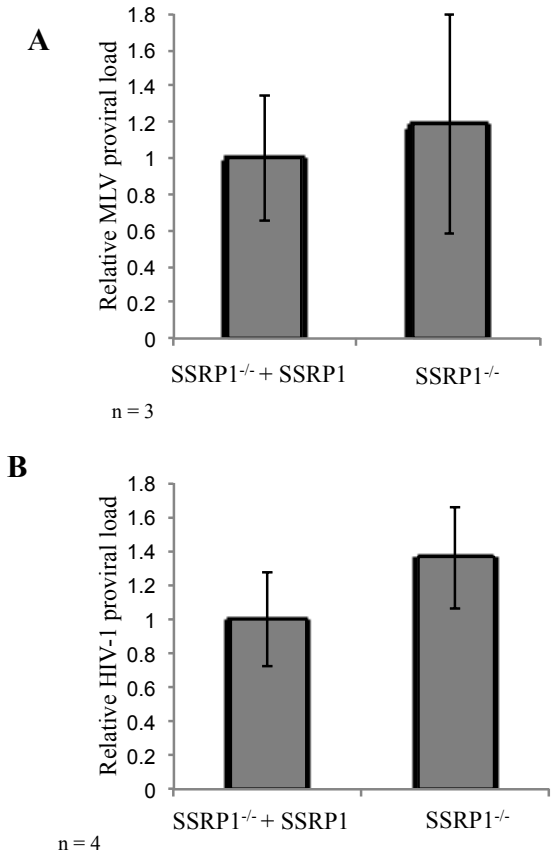



**Figure 3.7: The FACT complex does not promote gamma-retroviral or lentiviral integration.** (A) The FACT complex does not affect MLV integration. MLV proviral load was measured in infected cells by qPCR using viral specific primers relative to *GAPDH*. No significant difference in MLV integration frequency was observed in the *SSRP1* knockout cells (*SSRP1*$^{-/-}$) relative to cells expressing *SSRP1* (*SSRP1*$^{-/-}$ + *SSRP1*). (B) HIV-1 proviral load was not affected by varying abundance of SSRP1. HIV-1 integration frequency was measured by qPCR using viral specific primers relative to *GAPDH*. No significant difference in HIV integration frequency was observed in the *SSRP1* knockout cells (*SSRP1*$^{-/-}$) relative to cells expressing *SSRP1* (*SSRP1*$^{-/-}$ + *SSRP1*).

*The FACT complex has subtle effects on ALV integration pattern in vivo*

We hypothesized that the FACT complex might be targeting ALV integration due to the wide distribution of FACT complex throughout the genome, which is fitting with the observed random integration pattern for ALV. To test this we performed integration site mapping as previously described (Justice et al., 2015a) on our modified cell line (*SSRP1*$^{-/-}$ + *SSRP1)* or FACT knockout (*SSRP1*$^{-/-}$) cells. For each sample, we performed at least 2 biological replicates. We also made use of two distinct integration site mapping methods. The first is a sonication-based method that randomly fragments DNA for library preparation. The second method makes use of restriction digest to fragment the genome and is used to eliminate background from unintegrated viral products. Data from both methods agreed and thus, the data shown here is from the sonication-based library preparations due to reduced sequencing bias.

To preliminarily address whether the FACT complex might be targeting ALV integration in chicken cells, we analyzed integration site distribution using the HOMER bioinformatics toolkit. We recovered 13,277 and 12,692 unique integration sites for SSRP1$^{-/-}$ + SSRP1 and SSRP1$^{-/-}$ cell lines respectively. Looking at major genomic annotations such as transcription start sites (TSS), genes, promoters, CpG islands, and satellite sequences revealed that levels of *SSRP1* have small effects on integration distribution *in vivo* (Figure 3.8). The largest effect observed was the increased preference for integration into satellite sequences in cells lacking functional FACT complex ($p=0.005$).

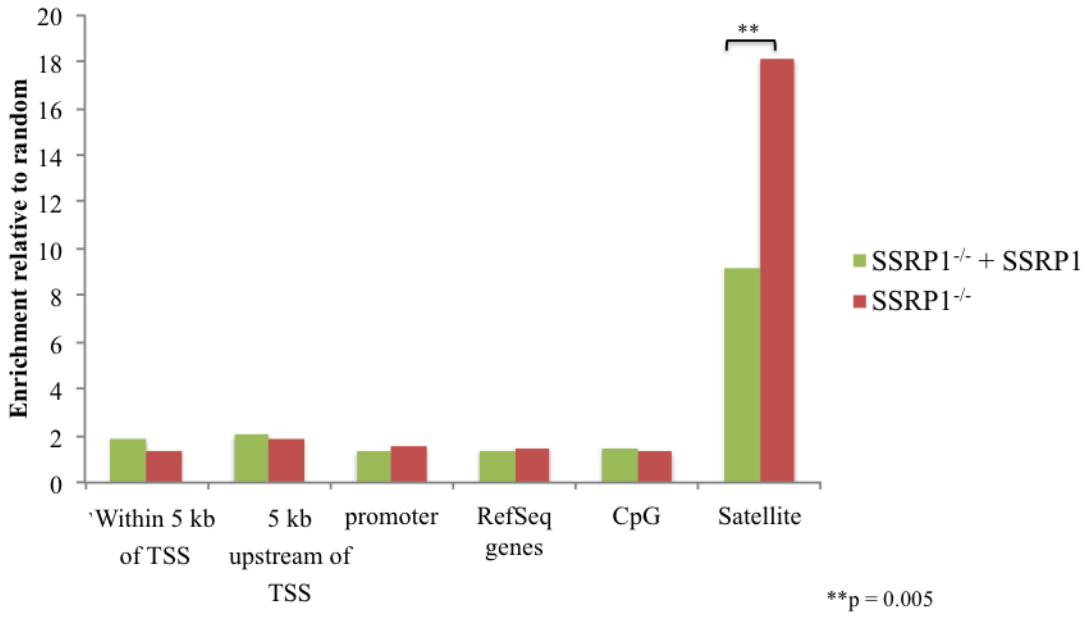



**Figure 3.8: Analysis of integration site pattern using HOMER bioinformatics program**. Shown is the enrichment of integration into specific genomic features relative to random. Significance was assessed using Fisher's exact test (**p<0.005)

Varying levels of FACT complex also had an effect on integration in the proximity of transcription start sites (Figure 3.9). We used HOMER bioinformatics tools to map each integration site to the nearest transcription start site (TSS). We then analyzed the 10 kb region flanking the TSS and plotted proportion of genes as a function of location relative to the TSS. We observed that when functional FACT complex is depleted, there is a significant decrease in integration in the proximity of the TSS ($p=0.0004$) Because the FACT complex plays a role in facilitating transcription it often co-localizes with RNA polymerase II, which has been shown to be poised near the TSS (Core and Lis, 2008). Thus, our evidence supports a role for the FACT complex in targeting ALV integration near the TSS.

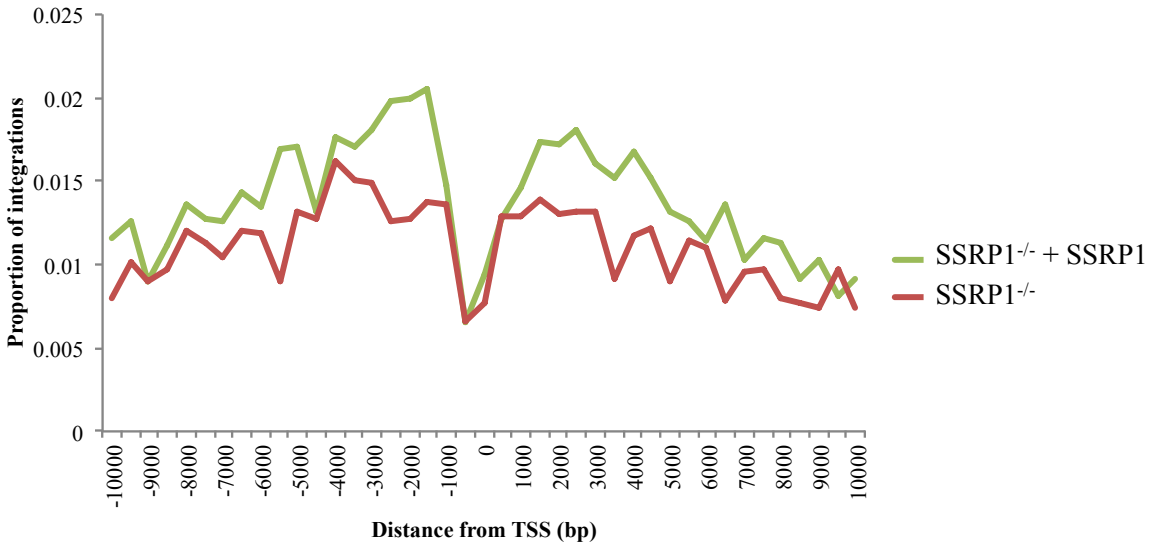
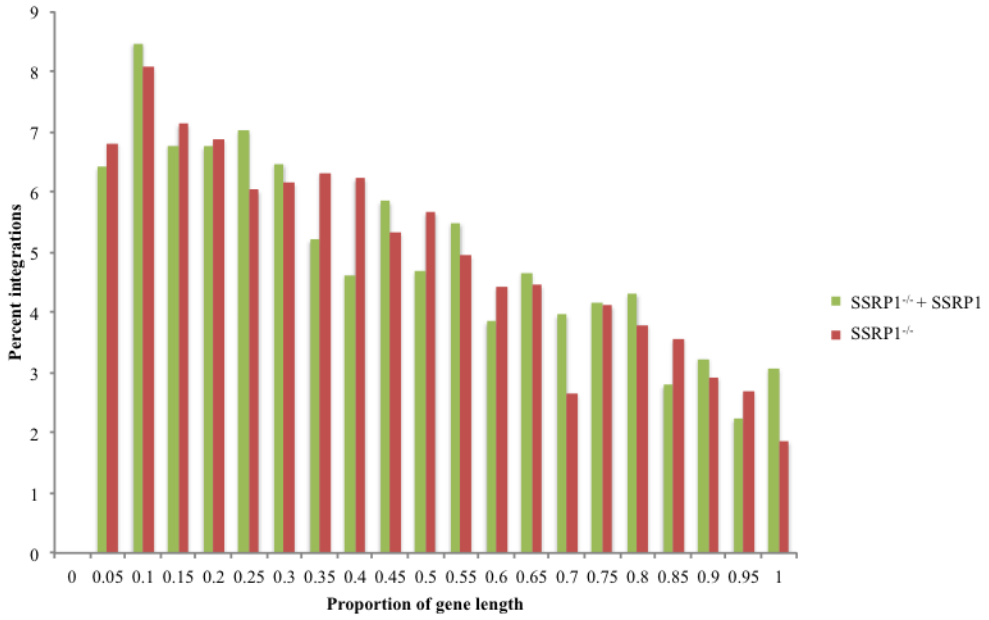
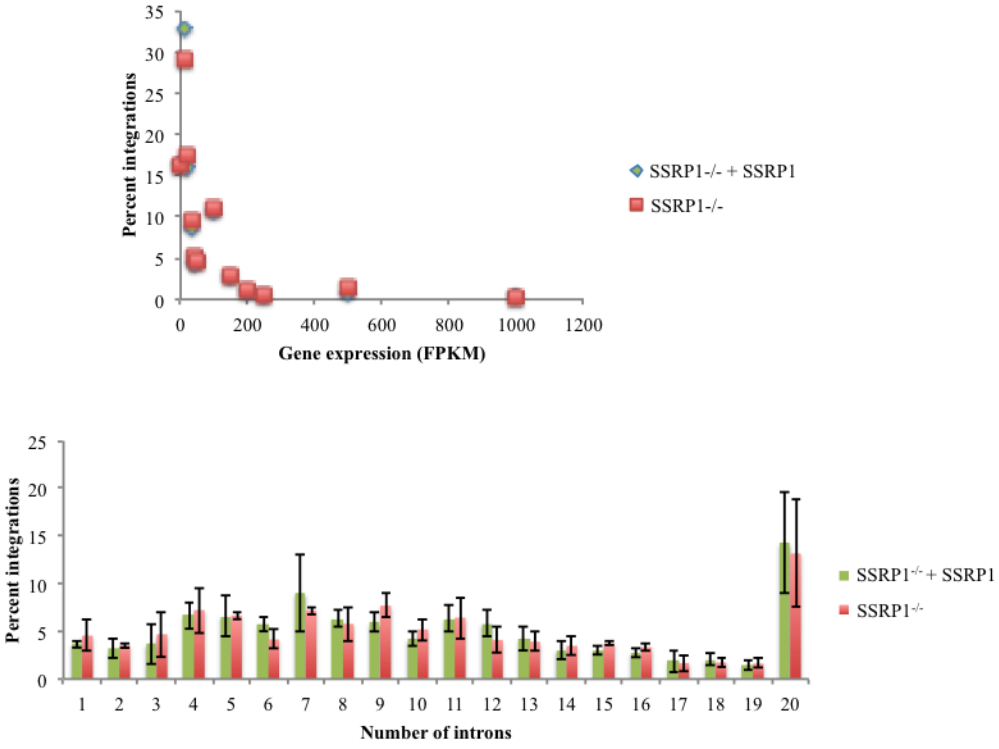



**Figure 3.9: Integration of ALV relative to TSS in WT and SSRP1 knockout cell lines**. Integration frequency is plotted as a function of distance from the transcription start site where the TSS is set to 0. Shown are the integration frequencies for cells with functional fact complex (SSRP1$^{-/-}$ + SSRP1) and those without (SSRP1$^{-/-}$). ROC analysis was used to assess significance (p=0.0004)

We find that the presence or absence of functional FACT complex had no effect on integration location within the gene body (Figure 3.10). We also correlated integration with gene expression levels and extent of splicing. We saw no significant differences between cells with functional FACT complex and the knockout cells in any category (Figure 3.11). Lastly, the sequence preference at the site of integration was not appreciably affected by the FACT knockout (Figure 3.12). Integration pattern in the SSRP1$^{-/-}$ and the SSRP1$^{-/-}$ + SSRP1 cell lines agreed with the observed distribution of integrations in DT40 cells (Chapter 2).




**<u>Figure 3.10</u>: Integration location throughout the gene body is not altered by SSRP1 knockout**. Of integrations that fell within a transcription unit, we analyzed where integrations were located throughout the gene body. To do this, we divided the gene body into 20 bins and plotted percent integrations that fell within each bin. A total of 2484 and 2425 integrations mapped within gene bodies for SSRP1$^{-/-}$ + SSRP1 and SSRP1$^{-/-}$ respectively. Data for cells with functional fact complex (SSRP1$^{-/-}$ + SSRP1) are shown in green and those without (SSRP1$^{-/-}$) are shown in red.




**<u>Figure 3.11</u>: ALV integration into expressed and spliced genes is unaffected by knockout of the FACT complex.** (A) To determine if the FACT complex affected integration into expressed genes, we analyzed the expression of the gene most proximal to the site of integration. We binned expression into 11 bins based on FPKM values and plotted percent of integrations that fell into each expression bin. Data for cells with functional fact complex (SSRP1$^{-/-}$ + SSRP1) are shown in blue and those without (SSRP1$^{-/-}$) are shown in red. (B) The extent of splicing of the most proximal gene to the site of integration was analyzed by plotting the percent of integrations that occur near genes based on the number of introns in that gene. Data for SSRP1$^{-/-}$ + SSRP1 is shown in green and SSRP1$^{-/-}$ is shown in red

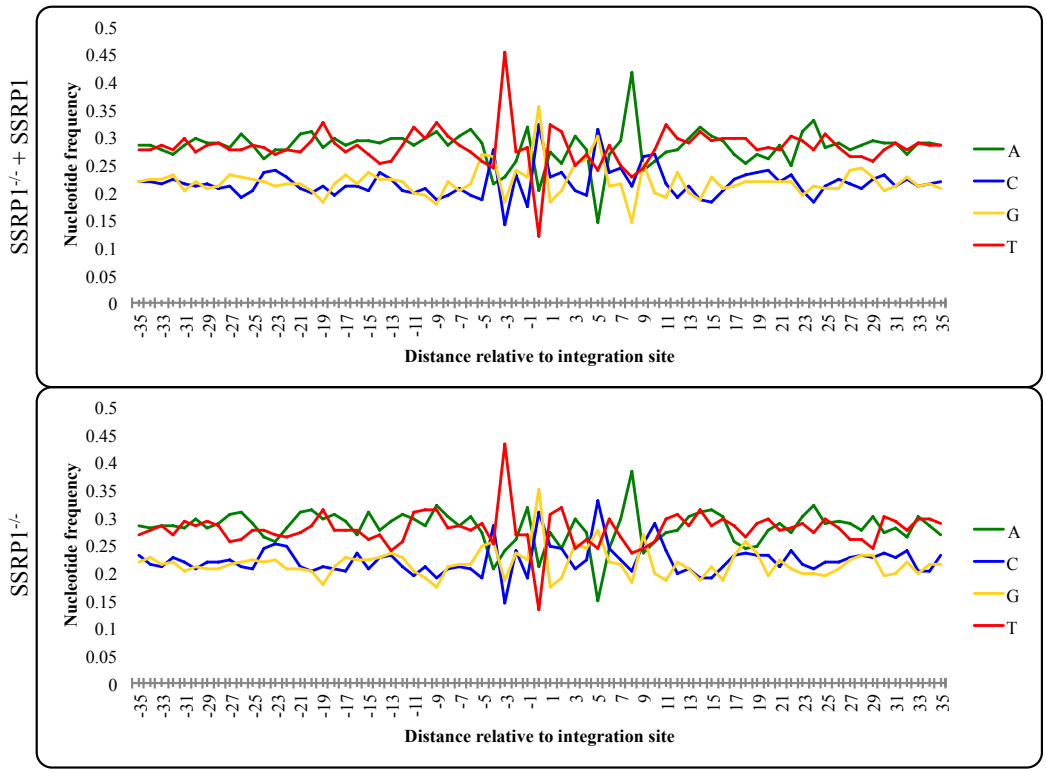
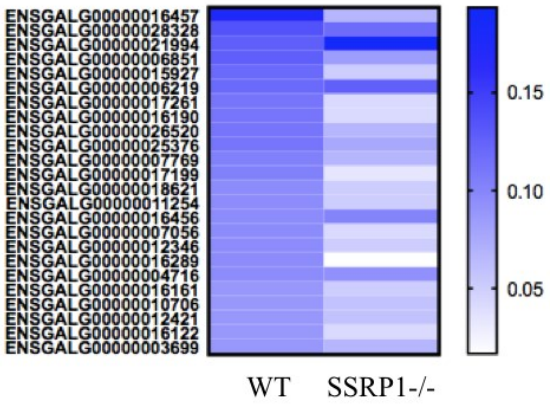



**Figure 3.12: Integration site sequence preference is not altered by levels of the FACT complex**. Nucleotide frequency at each base surrounding the site of integration is shown for cells expressing functional FACT complex and cells in which FACT complex has been depleted. The site of integration is set at 0.

*Common sites of integration differ between cells with and without functional FACT complex*

Certain genes are hotspots of integration and accrue multiple integrations. We analyzed recurrent integration genes (RIGs) between cells with and without functional FACT complex. RIGs in wild type cells seemed to be on average less targeted in FACT knockout cells (Figure 3.13A). To determine if the RIGs for each condition were enriched in any particular cellular pathways, we performed gene ontology analysis of all RIGs (defined as genes with 2+ integrations). Interestingly, the most enriched GO terms were all transcription factor binding sites and enriched transcription factor binding sites differed between conditions (Figure 3.13B) supporting the hypothesis that the FACT complex might be targeting ALV integration.

A




B

| WT | SSRP1$^{-/-}$ |
|---|---|
| C/EBP | Oct2 |
| Freac-3 | FOXD3 |
| GATA-3 | HNF1 |
| GATA-4 | miR-9 |
| GATA-5 | miR-516a |
| HOXA13 | miR-654 |
| miR-562 | POU1F1 |
| NKX2B | |
| TBP | |

**Figure 3.13: RIGs and GO term enrichment of RIGs in wild type vs. FACT knockout cells.** Integration sites were mapped to the nearest gene. Number of integrations per gene was calculated and RIGs were identified as genes with 2+ integrations. (A) Shown is a heatmap of percent of integrations into individual genes in wild type vs. FACT knockout cells. The top 25 most commonly targeted genes in wild type cells were analyzed. (B) All RIGs were used to analyze GO term enrichment. Genes were entered using the ordered query feature of gProfiler which weights genes according to how many integrations were detected. Shown are significantly ($p<0.05$) enriched transcription factor binding sites.

## 3.3 Discussion

In this study, we have identified the FACT protein complex, which is comprised of SSRP1 and Spt16, as the principal cellular binding partner of ALV IN. While ALV IN interacts with both SSRP1 and Spt16 individually, the FACT complex as a whole is required to stimulate integration activity *in vitro*. Additionally, we have demonstrated the importance of ALV IN CTD for binding to the FACT complex. Furthermore, we show that the level of ALV integration positively correlates with the abundance of the FACT complex in infected cells. The levels of 2-LTR circles were elevated when the FACT complex was depleted, demonstrating that this complex is critical for the integration step of the ALV life cycle. The level of plus strand extension products was unaffected by the levels of the FACT complex, indicating that the FACT complex does not affect reverse transcription. Taken together, our results elucidate a key role of the FACT complex in promoting ALV integration in infected cells. This regulation is likely not species-specific as interactions of ALV IN and the FACT complex were detected in both human and chicken cells. This is similar to what is seen for other retroviruses such as MLV IN which interacts with BET proteins from human and murine cells due to the high degree of conservation in chromatin binding proteins (Sharma et al., 2013).

Our findings indicate that the regulation of integration by the FACT complex is specific to alpharetroviral ALV. Unlike ALV IN, HIV-1 and MLV INs failed to bind the FACT complex. Moreover, MLV and HIV-1 integration was not significantly affected by altering the cellular levels of the FACT complex in infected cells.

HIV-1 and MLV exhibit strong integration site preferences for actively transcribed regions and gene regulatory regions such as transcription start sites and

enhancers, respectively (Lewinski et al., 2006; Schröder et al., 2002; Wu et al., 2003). In sharp contrast, ALV integration is relatively random and does not seem to target such regions as strongly (Chapter 2; Barr et al., 2005; Mitchell et al., 2004; Narezkina et al., 2004; Withers-Ward et al., 1994). Previous research has shown that the distinct integration site preferences of retroviruses can be linked to interaction of the virally encoded IN protein with various host cell factors (Craigie and Bushman, 2014; Debyser et al., 2015; Kvaratskhelia et al., 2014). Mechanistically, these host cell factors act largely as a bimodal tether to recruit the pre-integration complex to the chromatin thereby targeting proviral integrations (Kvaratskhelia et al., 2014). For example, LEDGF/p75 engages HIV-1 IN through its C-terminal integrase binding domain and guides HIV-1 integration to active genes (Cherepanov et al., 2003, 2005; Ciuffi et al., 2005; Llano et al., 2006; Maertens et al., 2003). The selection of the chromatin sites for integration is affected by the preferential binding of the N-terminal PWWP domain of LEDGF/p75 with the H3 histone tail containing trimethylated Lys36 (H3K36me3) a hallmark of actively transcribed genes (Eidahl et al., 2013). In a very similar manner, BET proteins, with their dual bromodomains, are able to guide MLV integration by bimodal interaction with both MLV IN and acetylated histone marks (Crowe et al., 2016; Larue et al., 2014; Sharma et al., 2013).

The FACT complex is believed to destabilize the histone octamer providing access to the chromosomal DNA for various enzymes (Formosa; Reinberg and Sims, 2006; Winkler and Luger, 2011). In particular, the C-terminal tail of Spt16 displaces nucleosomal DNA, allowing for access to the histone octamer (Winkler et al., 2011). This

capability to make chromatin more accessible by loosening or releasing the chromosomal DNA could allow for more effective integration.

In fact, a recent report shows that the chromatin states generated by overexpression or knockout of the FACT complex promote HIV-1 integration (Matysiak et al., 2017). This activation of HIV-1 integration was found to be dependent on the presence of nucleosomes and the chromatin remodeling activity of the FACT complex. The activation was recapitulated when chromatin states similar to those generated by FACT complex activity were artificially induced indicating that it is indeed the structures generated that promote HIV-1 integration. This mechanism differs from what we observe here for ALV integration. The FACT complex directly binds ALV integrase and integration is clearly correlated with FACT complex abundance. This suggests that the FACT complex may be acting as a tether to recruit the pre-integration complex to the chromatin similar to what is observed for LEDGF and HIV-1.

The evidence presented here strongly supports a role for the FACT complex in promoting ALV integration *in vitro* and *in vivo*. We see subtle effects of FACT knockout on ALV integration pattern, such as a depletion of integrations around the TSS and enrichment for integrations into satellite sequences. The depletion of integration in proximity to the TSS is consistent with a role for the FACT complex in tethering the PIC to predicted FACT binding sites thereby influencing integration nearby. Unfortunately, FACT binding data is not available for chickens and thus, further work is necessary to conclusively show that the FACT complex is indeed acting as a tether. The enrichment of integrations in satellite sequences in the absence of FACT complex is interesting. While we do not know the nature of the satellite sequences targeted by ALV, satellite sequences

in general are often found in heterochromatic regions. I hypothesize that the FACT complex may normally be targeting integration to euchromatic regions where it would be localized, and in the absence of FACT, integration is redirected to heterochromatin. This could be due to an innate preference for integration into heterochromatin or the result of a secondary targeting factor that takes over in the absence of the FACT complex. Further studies are necessary to elucidate the exact mechanism for regulation of ALV integration by the FACT complex, which could in turn facilitate the development of ALV-based vectors for their potential application in human gene therapy.

# Chapter 4 – Other host cell factors that regulate ALV replication: BET proteins, Nucleolin and UBTF

**Summary**

In addition to the two components of the heterodimeric FACT complex, Brd2, NCL (nucleolin) and UBTF (upstream binding transcription factor) were identified as specific binding factors of the ALV integrase protein in an affinity coupled mass spectrometry screen. Brd2 is a member of the BET family of proteins known to be important for regulating MLV integration. NCL has multiple functions including histone chaperone activity similar to that of the FACT complex. UBTF is a RNA polymerase I specific factor. Here we show that while NCL and UBTF do not directly affect integration of ALV, BET proteins seem to regulate ALV integration in competition with the FACT complex. To support this we find a role for BET proteins in inhibiting integration efficiency, but this effect is superseded by the effect of the FACT knockout. Further, while BET protein inhibition detectably affects integration targeting alone, inhibition in the context of a FACT knockdown has a larger effect on integration pattern. These data combined suggest that BET proteins may play a secondary role in both regulating and targeting ALV integration.

## 4.1 Introduction

Given that the FACT complex only had subtle effects on integration targeting, we hypothesized that additional host cell factors may be regulating ALV integration in addition to the FACT complex. While the FACT complex proteins (SSRP1 and SPT16) were the most promising hits to emerge from our screen for ALV integrase binding partners, other candidates that specifically bound ALV integrase were detected (Table 3.1). Other candidate host cell factors include Brd2, NCL (nucleolin) and UBTF (upstream binding transcription factor).

Brd2 is a member of the BET (bromodomain and extra terminal) family of proteins that includes Brd2, Brd3 and Brd4. Through their bromodomains, BET proteins interact with acetylated lysines of histone tails, specifically those of the H4 histone (Florence and Faller, 2001). BET proteins have known roles in regulating transcriptional elongation, cell cycle regulation and cancer (Florence and Faller, 2001). The BET family of proteins has also previously been shown to target MLV integration to transcription start sites and CpG islands by acting as a bimodal tether (Larue et al., 2014; De Rijck et al., 2013; Sharma et al., 2013). Brd4 specifically has been shown to antagonize transcription of HIV-1 (Bisgrove et al., 2007; Zhu et al., 2012a). BET proteins have also been shown to regulate the replication of many other viruses in various ways including the human herpesviruses, Epstein Barr virus (EBV) and Kaposi's sarcoma herpesvirus (KSHV), as well as papillomaviruses (Ilves et al., 2006; Lin et al., 2008; You et al., 2006).

Nucleolin is a multifaceted protein with various functions. Nucleolin was initially discovered as nucleolar factor important for RNA polymerase I transcription as well as

rRNA processing (Ginisty et al., 1998; Rickards et al., 2007). However, it has since been shown to also be found at the cellular membrane where it plays a role in mediating viral entry into host cells for various viruses such as human parainfluenza virus and respiratory syncytial virus (Bose et al., 2004; Tayyari et al., 2011). Interestingly, NCL also directly binds and regulates the RNA-dependent RNA polymerase of Hepatitis C virus (HCV) (Hirano et al., 2003). NCL has also been shown to regulate capsid assembly of adeno-associated virus (AAV) and be important for the episome maintenance functions of the EBNA1 protein of Epstein-Barr virus (Chen et al., 2014; Qiu and Brown, 1999). Nucleolin (NCL) is a particularly interesting candidate in our screen due to its functional similarity to the FACT complex (Angelov et al., 2006; Mongelard and Bouvet, 2007). In fact, the only proteins ever described to stimulate transcription from a chromatin template *in vitro* are the FACT complex, nucleolin and Brd2 (Angelov et al., 2006; LeRoy et al., 2008; Orphanides et al., 1998). That all of these proteins were found to specifically bind ALV integrase suggests that there may be a critical shared feature that the integrase protein binds or a common function that promotes integration.

Lastly, UBTF (upstream binding transcription factor) is a required component of the RNA polymerase I initiation complex (Kwon and Green, 1994). It has also been implicated in pre-ribosomal RNA processing as well as chromatin remodeling (Sanij et al., 2015). It has been described as an inhibitor of herpes simplex virus replication *in vivo* (Ouellet Lavallée and Pearson, 2015).

Here we find that BET proteins alone have a negative effect on ALV integration efficiency *in vivo* and subtle effects on ALV integration targeting. Interestingly, there seems to be competition between the FACT complex and BET proteins. In the presence

of the FACT complex, BET protein inhibition promotes ALV integration, however in the absence of the FACT complex, BET inhibition inhibits ALV integration efficiency. The fact that the FACT knockout phenotype is epistatic to the BET inhibition phenotype indicates that the FACT complex is likely the main host factor that regulates ALV integration and BET may act as a secondary factor. It seems though that while the FACT complex normally promotes ALV integration, BET proteins may inhibit the process. Inhibition of BET protein function has a stronger effect on ALV integration pattern when the FACT complex is absent further indicating that FACT may be the primary cofactor of ALV while BET proteins play a less important role. We further find that the other candidate host factors, NCL and UBTF, do not directly affect ALV integration efficiency *in vivo*.

## 4.2 Results

*BET protein inhibition affects ALV integration efficiency in vivo*

Brd2 was the 3rd most enriched protein in the ALV integrase pulldown fraction in our original mass spectrometry screen. Brd2 belongs to the BET family of proteins which include Brd2, Brd3 and Brd4. JQ1 is a small molecule inhibitor of BET protein binding to the chromatin (Filippakopoulos et al., 2010) . In order to assess whether BET proteins regulate ALV integration, we treated cells with JQ1 to inhibit BET protein function and infected with either ALV or MLV as a positive control (Figure 4.1). As expected, we see that MLV integration decreases by approximately 2-fold ($p<0.005$) when BET protein function is inhibited in agreement with previous publications (Aiyer et al., 2014; De Rijck et al., 2013; Sharma et al., 2013). However, we observed that in the absence of BET protein function, ALV integration efficiency actually significantly increased by slightly more than 2-fold ($p<0.05$).

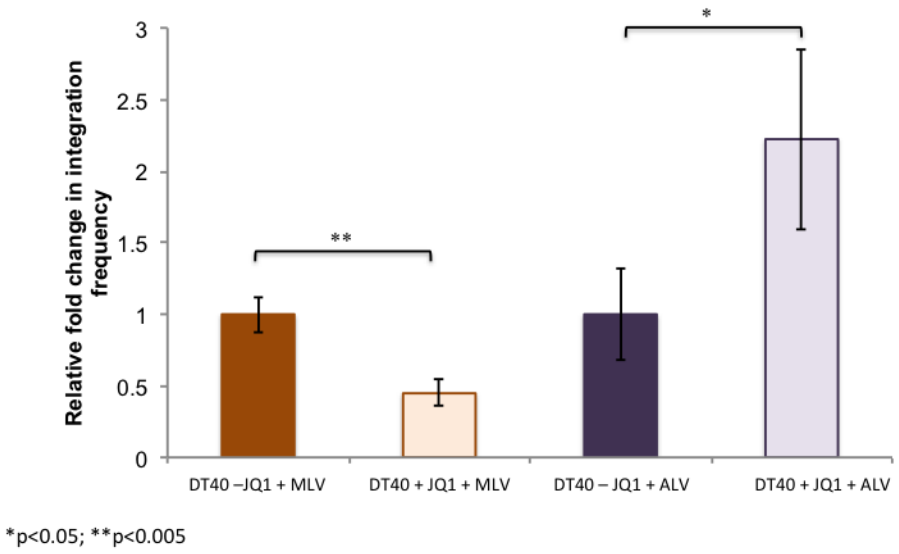



**Figure 4.1: BET protein inhibition promotes ALV integration efficiency in vivo**. DT40 cells were treated with JQ1, a small molecule inhibitor of BET protein function. Cells were then infected with either MLV, as a positive control, or ALV. Integration efficiency was measured using CR1-gag nested PCR approach and normalized to GAPDH. Shown are the fold changes in integration frequency after JQ1 treatment relative to untreated.

To determine if the observed effect of BET protein inhibition on ALV integration efficiency was at the level of integration we measured other viral intermediates such as reverse transcription (RT) products and 2-LTR circles. There was no detectable difference in RT product abundance, indicating that entry and RT occur comparably in the presence or absence of functional BET proteins (data not shown). We next looked at 2-LTR circle abundance. 2-LTR circles can serve as a marker for nuclear import as well as failed integration. If nuclear import is compromised, 2-LTR circle levels decline. On the other hand, if integration is blocked, 2-LTR circle abundance increases (Butler, Hansen, and Bushman 2001; Mandal and Prasad 2009). As expected in the case of MLV, where BET proteins are known to directly promote integration, JQ1 treatment led to a significant increase in 2-LTR circle levels (Figure 4.2). Interestingly, ALV 2-LTR circle abundance decreases in the presence of JQ1. This could indicate that nuclear import is compromised, but the lower 2-LTR circle levels in conjunction with higher integration frequency likely indicates that integration is directly being promoted when functional BET proteins are absent.

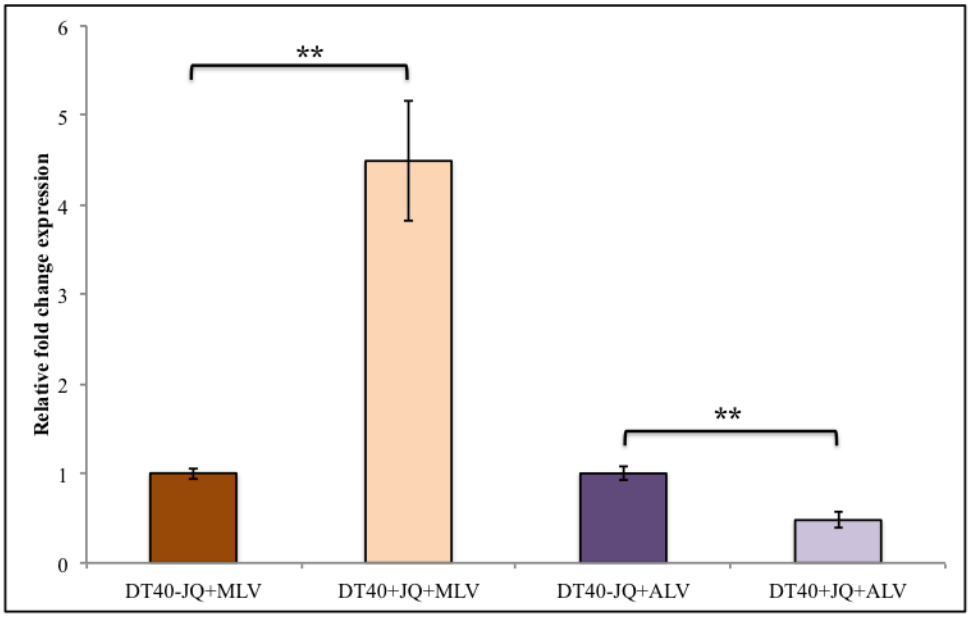



**Figure 4.2: BET protein inhibition causes a decrease in ALV 2-LTR circle levels.** DT40 cells were treated with JQ1 to inhibit BET protein infection and infected with either MLV or ALV. 2-LTR circle abundance was measured using qPCR and normalized to GAPDH. Shown is the fold change in abundance in JQ1 treated cells vs. untreated cells (**$p < 0.005$)

*BET protein inhibition has subtle effects on the ALV integration pattern in vivo*

Our evidence suggests that BET proteins play a role in inhibiting ALV integration, opposite to the effect of the FACT complex. We next analyzed whether this effect on ALV integration frequency was accompanied by an effect on targeting. To determine this we mapped ALV integration sites in WT DT40 cells in the presence or absence of the BET protein small molecular inhibitor JQ1. We specifically chose to look at transcription start sites, promoters, genes and CpG islands because these are known BET protein binding sites. The inhibition of BET proteins in a wild type background had little effect on integration into any of these features (Table 4.1). The most notable difference was integration into genes, which changed by roughly 4% ($p < 0.1$). Upstream of the TSS, in the promoter region, BET protein inhibition also decreased integration ($p < 0.1$). Integrations in the transcription start sites and CpG islands were not significantly affected by JQ1 treatment.

| **Annotation** | **WT** | **WT+JQ1** |
|---|---|---|
| within 5kb of TSS | 10.5 | 9.0 |
| 5 kb upstream of TSS | 7.6 | 5.9 |
| RefSeq genes | 37.9 | 41.6 |
| CpG-Island | 3.8 | 2.8 |

**Table 4.1: Integration frequency in common genomic features in wild type and JQ1-treated wild type cells.** Percent of detected ALV integrations that fell within each specified category was determined using HOMER bioinformatics tools. Significant differences between WT and JQ1 treated cells was assessed using Fisher's exact test.

*FACT complex and BET proteins seem to have competing effects on ALV integration*

We next wanted to examine whether the FACT complex and BET proteins had competing effects on ALV integration *in vivo*. In order to do this, we inhibited BET protein function in the presence of varying levels of FACT complex (Figure 4.3). In wild type cells and FACT complex overexpressing cells, BET protein inhibition promoted ALV integration efficiency. Interestingly, the more FACT complex present, the larger the effect of BET protein inhibition was. However, when BET protein function was inhibited in a FACT knockout background, integration efficiency actually decreased. This phenotype is comparable to what is seen with the FACT knockout alone, indicating that the effect of the FACT complex on ALV integration is epistatic to that of the BET proteins.

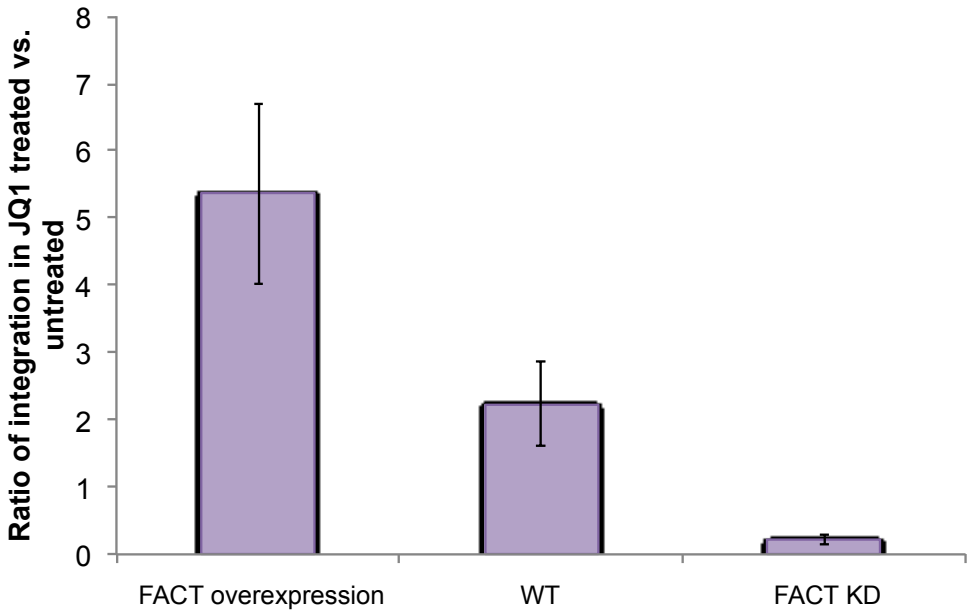



**Figure 4.3: Effect of BET protein inhibition on ALV integration efficiency in the presence of varying levels of the FACT complex**. Cells with either overexpressed, wild type or knockout levels of FACT complex were treated with JQ1 to inhibit BET protein function. ALV integration efficiency was then measured using CR1-gag nested qPCR approach and normalized to GAPDH. Plotted is the ratio of integration observed in JQ1 treated cells from each condition relative to the matched untreated condition (**$p<0.005$).

Since our data suggested that BET proteins may play a less important role in regulating ALV integration as compared to the FACT complex, we reasoned that perhaps BET protein inhibition might more appreciably affect ALV integration pattern in the absence of the FACT complex, the primary host cell factor. Thus, we mapped integration sites in FACT knockout cells or JQ1-treated FACT knockout cells. Percent of integrations into selected genomic features is shown in Table 4.2. We observe a nearly 2-fold increase in integrations near the TSS and a 14% increase in integrations into RefSeq genes.

| Annotation | FACT KO | FACT KO + JQ1 |
|---|---|---|
| within 5kb of TSS | 4.7* | 8.3* |
| 5 kb upstream of TSS | 3.3 | 4.0 |
| RefSeq genes | 35.8* | 40.7* |
| CpG-Island | 4.8 | 3.8 |

**Table 4.2: Integration frequency of ALV into various genomic features in FACT knockout and FACT knockout JQ1 treated cells**. Percent of integrations into the listed genomic features was calculated using HOMER bioinformatics tools. Significant differences between conditions was assessed using Fisher's exact test (*$p<0.05$)

However, to further interrogate any potential differences, we took a closer look at integrations in the proximity of transcription start sites and CpG islands. We did this for all four of our tested conditions (wild type, wild type JQ1 treated, FACT knockout and FACT knockout JQ1 treated). There was no difference in integration in the proximity of these features between the wild type and JQ1-treated wild type cells (data not shown). However, there were striking differences in integration near both transcription start sites and CpG islands when FACT knockout cells were treated with JQ1 to inhibit BET protein function (Figure 4.4). We observe that when BET protein function is inhibited, there are more detectable integrations in the region flanking the transcription start site (Figure 4.4A) and less integration in the region surrounding CpG islands (Figure 4.4B). This seems to indicate that normally BET proteins may have two roles in ALV integration targeting - occluding ALV integration near the TSS and promoting integration near CpG islands. The fact that these effects are only evident in the absence of the FACT complex again points to a primary role for the FACT complex in regulating ALV integration.

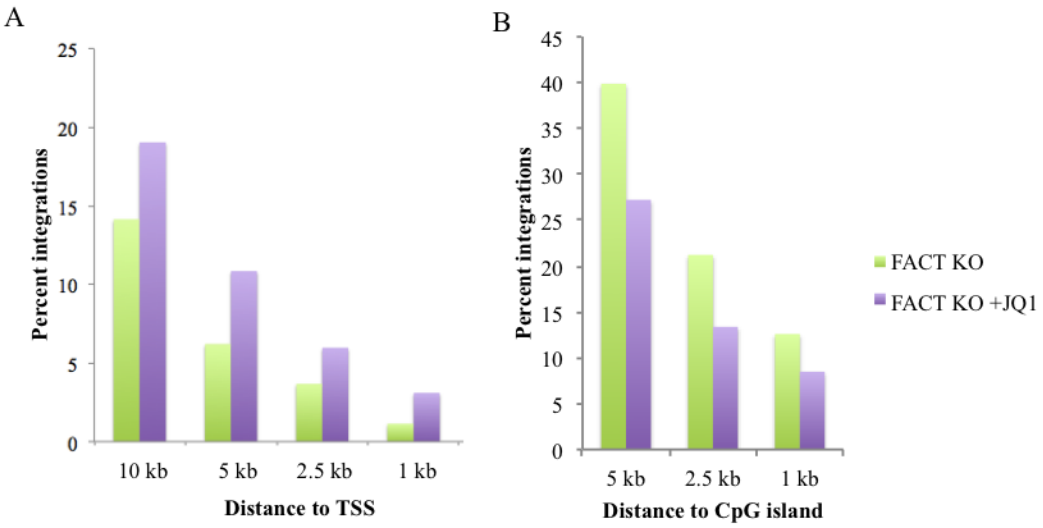



**Figure 4.4: Integration in the proximity of transcription start sites and CpG islands in FACT knockout and JQ1-treated FACT knockout cells**. Integration in the proximity of both features was calculated using BedWindow to determine the number of integrations that fell within a specified window surrounding the genomic feature. This was then normalized to total detected integrations. Shown is the percent integrations that fall into each specified category.

*UBTF affects early steps of ALV life cycle*

After Brd2, the 4$^{th}$ most enriched protein that bound specifically to ALV integrase was UBTF (upstream binding transcription factor). To determine the effect of UBTF on ALV replication, we made use of targeted siRNA against human UBTF to knock down expression levels in HEK293T cells (Figure 4.5A). A single siRNA was sufficient to reduce UBTF expression levels approximately 8-fold ($p<0.05$) relative to a scrambled siRNA control.

We then infected UBTF knockdown cells with pseudotyped ALV and measured ALV integration frequency (Figure 4.5B). We observed an approximate 2-fold reduction in proviral integration frequency ($p<0.05$). However, because integration is the final step in the early life cycle of retroviruses, any effect of the knockdown on earlier viral processes would also manifest as a reduction on proviral load. Thus, we measured plus strand extension products (PSE), a late product of reverse transcription (Figure 4.5C). We observed that in cells with depleted UBTF expression, there was a 7-fold decrease in PSE products ($p<0.05$). This indicates that the absence of UBTF has a significant effect on the process of reverse transcription. We also did not observe any differences in 2-LTR circle abundance between wild type and UBTF knockdown cells (data not shown). Taken together, these data indicate that UBTF may be playing a role in regulating reverse transcription but has no significant effects on the process of integration.

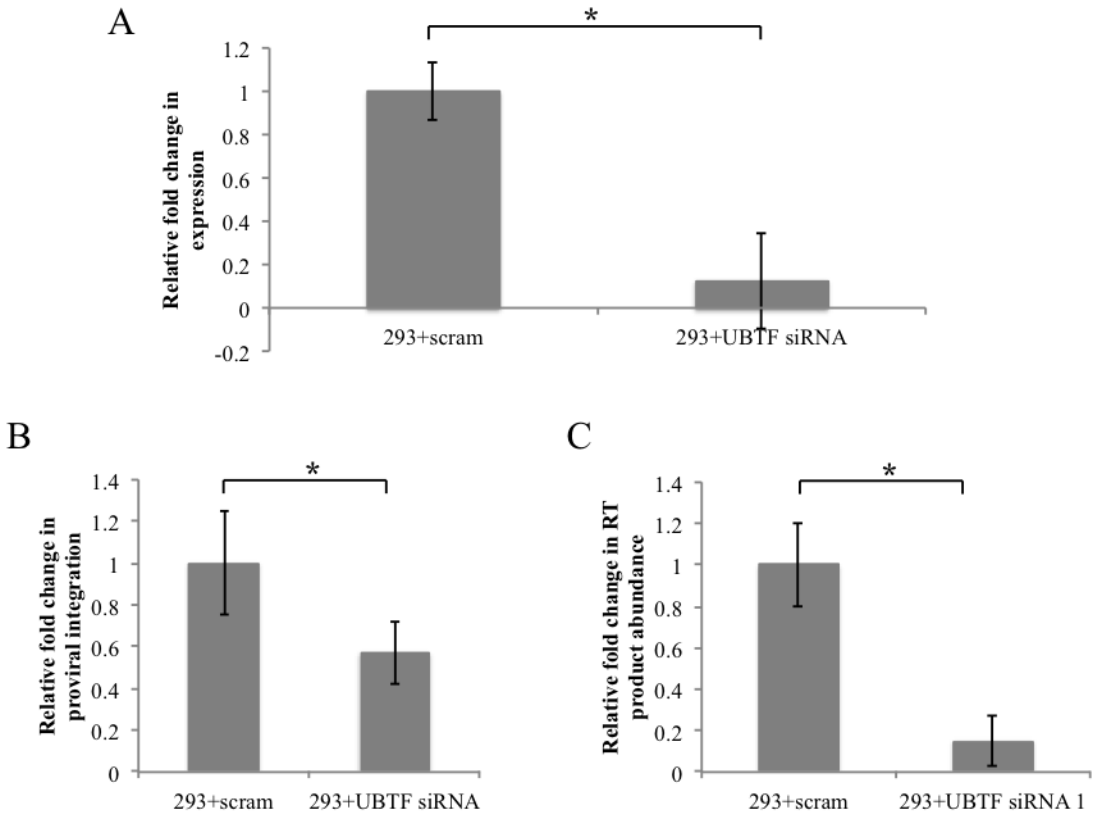



**Figure 4.5: UBTF affects ALV replication**. (A) siRNA mediated knockdown of UBTF. Expression levels of UBTF were measured in cells transfected with siRNA targeting either UBTF or a scrambled control using qPCR. Expression was normalized to a housekeeping gene, GAPDH (*$p<0.05$). (B) CR1-gag nested qPCR was used to measure proviral integration frequency. Proviral load was normalized to a housekeeping gene, GAPDH. Shown is the fold change in detectable integrants relative to the scrambled control (*$p<0.05$). (C) Plus strand extension products were measured using primers in gag, and normalized to a housekeeping gene, GAPDH (*$p<0.05$).

*NCL has no effect on ALV replication*

Nucleolin (NCL) was also significantly enriched in our screen for ALV integrase binding partners. NCL also has functional similarity to the FACT complex making it an intriguing candidate gene. To determine if NCL may be affecting ALV replication, we knocked down NCL expression levels in HEK293T cells using targeted siRNA (Figure 4.6A). Two individual siRNA both depleted NCL expression levels, one to essentially 0 and the other roughly 3-fold ($p<0.05$).

To determine affects of NCL on the ALV life cycle, we infected cells transfected with siRNA targeting either NCL or a scrambled control with pseudotyped ALV and measured ALV integration frequency (Figure 4.6B). We observed no significant differences in ALV proviral integration frequency when NCL was knocked down. Thus, NCL does not have an appreciable effect on ALV replication *in vivo*.

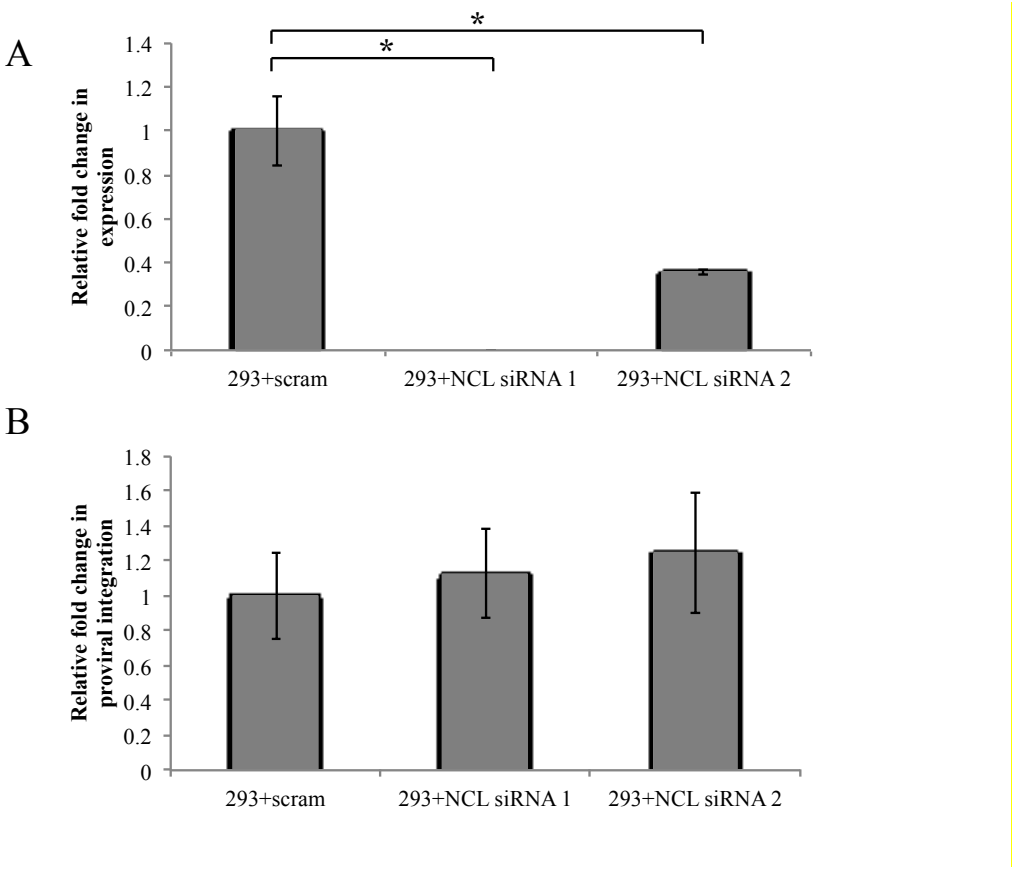



**Figure 4.6: NCL has no effect on ALV integration.** (A) Expression levels of NCL were measured in cells transfected with siRNA targeting NCL or a scrambled control using qPCR and normalized to GAPDH. (B) Proviral integration frequency was determined using the CR1-gag nested PCR approach. Integration frequency was normalized to GAPDH.

## 4.3 Discussion

We find that nucleolin (NCL) and UBTF have no direct role in regulating ALV integration. NCL knockout had no effects on ALV integration efficiency in cultured cells *in vitro*. It is possible that the presence of functional FACT complex masks any affects the NCL knockout may have had. However, this is difficult to test in our hands due to the fact that our established FACT knockout cell line system is constructed in B-cells which are unable to be efficiently transfected making siRNA mediated knockdown difficult. UBTF knockdown did have affects on ALV integration frequency in a wild type background, however it was determined based on reverse transcription (RT) product abundance that this was the result of compromised RT in the absence of UBTF.

Interestingly, inhibition of the BET proteins did seem to have an effect on ALV integration. In the presence of overexpressed FACT complex, BET protein inhibition increased ALV integration efficiency nearly 5-fold. In a wild type background, BET protein inhibition increased ALV integration frequency to a lesser extent, about 2-fold. In the absence of functional FACT complex however, ALV integration efficiency was decreased. We believe this indicates that the FACT complex and BET proteins have competing effects on ALV integration. BET proteins inhibit ALV integration while FACT complex promotes integration. In the case of overexpressed FACT, when BET protein inhibition of integration is relieved, abundant FACT promotes integration to a large extent. This effect is less pronounced in the wild type situation because there is less FACT complex present to promote integration. Finally, in the case of the FACT knockout, lifting the normal inhibition of ALV integration by the BET proteins has no effect because there is no FACT complex present to direct integration and thus the FACT

knockout phenotype is dominant. This seems to suggest that the FACT complex plays a more primary role in regulating ALV integration efficiency while effects of the BET proteins may be secondary. This is supported by the fact that ALV integration targeting is only appreciably affected by BET protein inhibition when functional FACT complex is absent.

This model needs to be further tested. An obvious way to do so would be to test BET protein binding to ALV integrase *in vitro*. If BET does indeed bind ALV IN, then competitive binding experiments with FACT complex components would be valuable. More integration site mapping data in chickens and humans would also be helpful. Our analyses were limited by the relatively small number of integration sites. Further, our analysis was performed in chicken cells, which limits the features we are able to analyze simply due to a poorly annotated reference genome. If these experiments were repeated in human cells for which BET protein and FACT complex binding data is available, proximity of integrations to these locations could be calculated and potentially reveal differences in targeting that we are unable to see in our system.

# Chapter 5 - Integration of ALV into *CTDSPL* and *CTDSPL2* genes in B-cell lymphomas promotes cell immortalization, migration and survival

Adapted from:
**Winans S**., Flynn A., Malhotra S., Balagopal V., Beemon K. (2017). Integration into *CTDSPL* and *CTDSPL2* genes in B-cell lymphomas promotes cell immortalization, migration and survival. *Oncotarget* 8(34):57302-57315.

**Summary**

Avian leukosis virus induces tumors in chickens by integrating into the genome and altering expression of nearby genes. Thus, ALV can be used as an insertional mutagenesis tool to identify novel genes involved in tumorigenesis. Deep sequencing analysis of viral integration sites has identified *CTDSPL* and *CTDSPL2* as common integration sites in ALV-induced B-cell lymphomas, suggesting a potential role in driving oncogenesis. We show that in tumors with integrations in these genes, the viral promoter is driving the expression of a truncated fusion transcript. Overexpression in cultured chick embryo fibroblasts reveals that *CTDSPL* and *CTDSPL2* have oncogenic properties, including promoting cell migration. We also show that *CTDSPL2* has a previously uncharacterized role in protecting cells from apoptosis induced by oxidative stress. Further, the truncated viral fusion transcripts of both *CTDSPL* and *CTDSPL2* promote immortalization in primary cell culture.

## 5.1 Introduction

Our lab has previously identified *CTDSPL* (C-terminal domain small phosphatase-like) and *CTDSPL2* (C-terminal domain small phosphatase-like 2) as common integration sites in ALV-induced B-cell lymphomas (Justice et al., 2015b). The recurrence and selection of integrations within these genes in tumors suggests that they may be involved in driving tumorigenesis. The CTDSP family of proteins consists of CTDSP1, CTDSP2, CTDSPL and CTDSPL2 proteins, all of which contain a catalytic FCP1 (F-cell production 1) homology domain that functions as a phosphatase (Yeo et al., 2003). The CTDSP family has been shown to dephosphorylate the C-terminal domain (CTD) of RNA polymerase II *in vitro* (Thompson et al., 2006; Yeo et al., 2003). Through this function, this family of proteins is proposed to be important for transcriptional regulation. Most family members preferentially dephosphorylate Ser5 of the CTD and thus control the transition from initiation to processive transcription elongation (Thompson et al., 2006; Yeo et al., 2003). CTDSP1, CTDSP2 and CTDSPL have also been shown to play a role in gene silencing, most notably of neuronal gene expression, through interaction with the REST complex (Thompson et al., 2006; Visvanathan et al., 2007; Yeo et al., 2005).

The CTDSP proteins are able to act on additional targets as well. For instance, CTDSP1/2/L proteins have been shown to induce TGF-β signaling and attenuate BMP signaling (Knockaert et al., 2006; Wrighton et al., 2006). CTDSP1 also stabilizes SNAIL and C-MYC proteins by dephosphorylating a key serine residue in each protein (Wang et al., 2016; Wu et al., 2009). Further, *CTDSP1/2/L* genes have all been found to contain an intronic microRNA that belongs to the miR-26 family. These miRNAs have been shown

to act synergistically with the CTDSP1 and CTDSPL proteins to dephosphorylate, and thus activate, pRb and block the G1/S cell cycle transition (Zhu et al., 2012b). CTDSP2 has also been shown to inhibit cell cycle progression independently by activating Ras and p21 (Kloet et al., 2015).

Due to involvement in these pathways, it comes as no surprise that the CTDSP1/2/L and the miR-26 family have been implicated in tumorigenesis. *CTDSPL* has been characterized as a tumor suppressor gene that is frequently deleted or mutated in many major epithelial cancers such as lung, renal cell and breast carcinoma (Kashuba et al., 2004, 2009; Senchenko et al., 2010; Zhu et al., 2012b). Further, all 3 proteins are down-regulated in hepatocellular carcinoma cell lines (Zhu et al., 2012b). Comparatively, little is known about CTDSPL2. It has been shown to play a role in erythroid differentiation and BMP signaling (Wani et al., 2016; Zhao et al., 2014). However, CTDSPL2 has not been previously linked to tumorigenesis.

In this work, we characterize *CTDSPL2* as a novel gene involved in oncogenesis and further characterize the role of *CTDSPL*. Specifically, we investigate the function of viral induced truncations of both genes in cancer. Overexpression of *CTDSPL* and *CTDSPL2* leads to changes in expression of ribosomal genes and genes involved in cellular migration and metabolism. We show that overexpression of both *CTDSPL* and *CTDPSL2* causes accelerated cellular migration in primary cell culture. Interestingly, expression of *CTDSPL2*, but not *CTDSPL*, protects cells from apoptosis induced by oxidative stress, indicating that the two genes may not be redundant. Importantly, the truncated viral fusion transcripts of both *CTDSPL* and *CTDSPL2* promote immortalization when overexpressed in primary cell culture.

## 5.2 Results

*CTDSPL and CTDSPL2 are common integration sites in ALV-induced B-cell lymphomas*

High throughput sequencing was used to identify retroviral integration sites in ALV-induced B-cell lymphomas (Justice et al., 2015b). Integration sites that are overrepresented in the sequencing data, either because of clonal expansion or because the gene is a common integration site between tumors, were selected for and therefore believed to be important in tumorigenesis.

*CTDSPL* and *CTDSPL2* were identified to be common integration sites previously (Justice et al., 2015b). In this study we have expanded our analysis and observed 23 unique clonally expanded integrations in *CTDSPL* in 12 tumors from 7 birds. All expanded integrations are in the same transcriptional orientation as *CTDSPL* and fall upstream of exon 4 (Figure 5.1; Table 5.1). In addition, thirteen unique expanded integrations were detected in *CTDSPL2* in 7 tumors from 4 birds. All expanded integrations in the gene are in the same transcriptional orientation as *CTDSPL2* and fall upstream of exon 3 (Figure 5.1; Table 5.1). No expanded integrations into either gene were observed in non-tumors. Interestingly, we did not observe integration into other CTDSP family members.

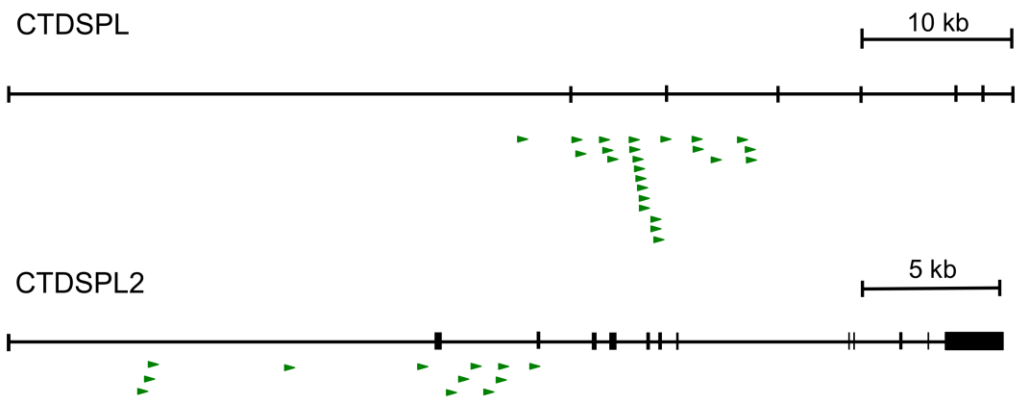



**Figure 5.1: *CTDSPL* and *CTDSPL2* are common integration sites in ALV-induced B-cell lymphomas**. A schematic of retroviral integrations into both *CTDSPL* and *CTDSPL2*. Each integration is depicted as an arrow with the base of the arrow representing the site of integration. Direction of the arrow indicates the orientation of the retroviral integration with respect to transcription of the gene; all are in the sense orientation. There are 23 unique expanded integrations in *CTDSPL*, all of which fall upstream of exon 4. There are 13 unique expanded integrations in *CTDSPL2* upstream of exon 3.

**Table 5.1: Genome coordinates, breakpoints and tumor information for integrations into CTDSPL and CTDSPL2**. All clonally expanded integrations (2 or more breakpoints) detected in the screen are listed with tumor ID, number of breakpoints and genomic coordinates that correspond to the site of integration. CTDSPL coordinates are on chromosome 2; CTDSPL2 coordinates are on chromosome 10.

| **CTDSPL** | | |
|---|---|---|
| Breakpoints | Tumor ID | Locus # |
| 71 | C3K | 4519921 |
| 53 | C3L | 4519921 |
| 51 | D2K | 4520683 |
| 26 | D2K | 4519498 |
| 25 | D2K | 4519227 |
| 24 | D2L | 4520683 |
| 22 | D2B | 4520683 |
| 18 | D5S | 4515471 |
| 16 | D2L | 4519227 |
| 14 | A1B | 4511874 |
| 13 | D5L | 4515471 |
| 12 | D2K | 4527016 |
| 10 | C2B | 4523438 |
| 6 | C2B | 4517279 |
| 5 | D2L | 4519498 |
| 5 | D2B | 4519498 |
| 5 | D2B | 4527016 |
| 4 | D2L | 4519620 |
| 4 | B6B | 4520674 |
| 4 | D2L | 4524658 |
| 4 | D2L | 4526898 |
| 4 | D2K | 4526898 |
| 4 | D2L | 4527016 |
| 3 | D2B | 4519344 |
| 3 | D2K | 4519620 |
| 3 | D2B | 4519675 |
| 3 | C3K | 4519923 |
| 2 | D5B | 4515471 |
| 2 | D9B | 4517448 |
| 2 | D5S | 4517790 |
| 2 | D2B | 4519227 |
| 2 | D2K | 4519827 |
| 2 | D2K | 4520864 |
| 2 | D2K | 4521356 |
| 2 | C3L | 4523454 |
| 2 | D2K | 4526408 |

| **CTDSPL2** | | |
|---|---|---|
| Breakpoints | Tumor ID | Locus # |
| 47 | D2K | 19283543 |
| 42 | D2K | 19283421 |
| 18 | D2B | 19283421 |
| 15 | D2B | 19283543 |
| 15 | D2L | 19283543 |
| 12 | D2L | 19283421 |
| 10 | D2K | 19296421 |
| 9 | D2K | 19282355 |
| 8 | C3L | 19284002 |
| 4 | D2L | 19282355 |
| 4 | C3L | 19285319 |
| 4 | C3K | 19296007 |
| 4 | D2B | 19296421 |
| 3 | D3K | 19286358 |
| 3 | D2L | 19296421 |
| 2 | D2B | 19282355 |
| 2 | D2B | 19284446 |
| 2 | D2K | 19284889 |
| 2 | C3K | 19285319 |
| 2 | D2K | 19291118 |
| 2 | C3L | 19296007 |
| 2 | D2B | 19296037 |
| 2 | A8B | 19310663 |

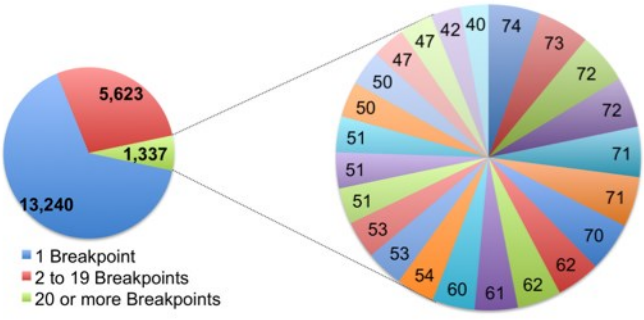
*Activation of CTDSPL and CTDSPL2 are likely early events in tumorigenesis*

A number of integration sites in the *CTDSPL* and *CTDSPL2* genes were found to be highly clonally expanded. Clonal expansion of a specific integration within a tumor was estimated via quantitation of sonication breakpoints as described previously (Justice et al., 2015b). The highest breakpoint integrations from tumors carrying *CTDSPL* and *CTDSPL2* integrations are shown in a composite pie chart in Figure 5.2A. In some individual tumors, these integrations were amongst the most dominant, expanded integrations (Figure 5.2B). This suggests that these integrations occurred early in tumorigenesis and were expanded as the tumor progressed.

Identical integration sites within both *CTDSPL* and *CTDSPL2* genes were identified in primary (bursal) and secondary (liver and kidney) tumors found in the same bird (Figure 5.2C). The presence of identical integration sites also indicates that these integrations likely occurred early in tumorigenesis prior to metastasis. The primary bursal tumor then metastasized to various locations including the liver, kidney and spleen, causing the clonal expansion of the integrated provirus in different secondary tumors. Interestingly, integrations in *CTDSPL* and *CTDSPL2* frequently occur in the same tumor. For instance, primary and secondary tumors in birds C3 and D2 have many of the most clonally expanded integrations in both genes (Table 5.1).

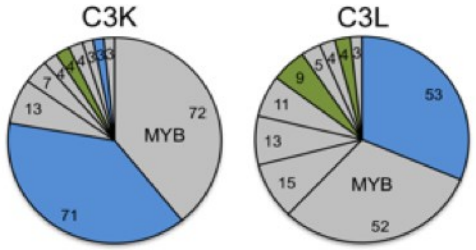
**Figure 5.2: Viral integrations into *CTDSPL* and *CTDSPL2* are an early event in tumorigenesis.** (A) A composite pie chart of 11 tumors containing clonally expanded integrations in either *CTDSPL* or *CTDSPL2* is shown. In this tumor set, there were 14,879 unique integrations represented by 20,200 total breakpoints. Integrations having 40 or more breakpoints are depicted in a separate composite pie chart with individual integrations as a single "slice" of the pie chart, weighted by number of breakpoints. There are 23 integrations with 40 or more breakpoints. Of these top clonally expanded integrations, 3 are in *CTDSPL* and 2 are in *CTDSPL2*. (B) *CTDSPL* and *CTDSPL2* are dominant integrations in some individual tumors. The top 10 clonally expanded integrations are shown for representative tumors (C3L and C3K). In these cases *CTDSPL* and *CTDSPL2* are among the most dominant integrations. *CTDSPL* integrations are indicated in blue, *CTDSPL2* integrations are indicated in green. Top integrations are labeled. (C) Primary (bursa) and secondary (kidney and liver) tumors from the same bird (D2) have identical integrations in *CTDSPL* and *CTDSPL2*. *CTDSPL* integrations are indicated in blue, *CTDSPL2* integrations are indicated in green. The most dominant integrations are labeled. Identical integrations in each tumor are indicated with *. These integrations are clonally expanded in all cases, comprising a comparable proportion of the total tumor breakpoints. This suggests that these integrations occurred early in tumorigenesis, within the bursa and subsequently metastasized to these secondary sites.

A

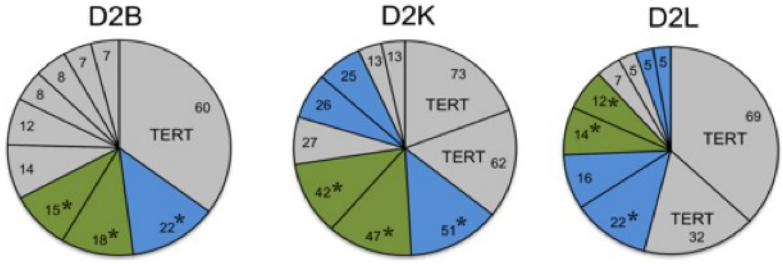



| Gene | Tumor ID | #Breakpoints |
|---|---|---|
| TERT | D2L | 74 |
| TERT | D2K | 73 |
| TERT | D2B | 72 |
| C-MYB | C3K | 72 |
| AHI1 | D5B | 71 |
| CTDSPL | C3K | 71 |
| AHI1 | D5S | 70 |
| NR3C1 | D5S | 62 |
| TERT | D2K | 62 |
| Unannotated | D5B | 61 |
| TERT | D2B | 60 |
| TAPBP | A1B | 54 |
| C-MYB | A1B | 53 |
| CTDSPL | C3L | 53 |
| AHI1 | D5L | 51 |
| CTDSPL | D2K | 51 |
| C-MYB | C3L | 51 |
| TERT | A1B | 50 |
| NR3C1 | D5B | 50 |
| NR3C1 | D5L | 47 |
| CTDSPL2 | D2K | 47 |
| CTDSPL2 | D2K | 42 |
| FAM19A2 | D5S | 40 |

B



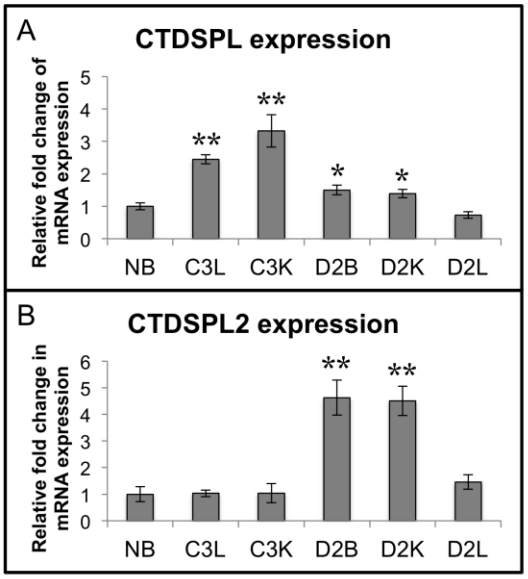

C




3

*Viral integrations in CTDSPL and CTDSPL2 drive the overexpression of genes.*

Quantitative RT-PCR verified that relative to normal bursa, levels of both *CTDSPL* and *CTDSPL2* mRNA were elevated in the tumors with highly clonally expanded integrations in these genes (Figure 5.3A). For instance, C3L and C3K had a co-dominant integration in *CTDSPL* (Figure 5.2B) and expression of this gene was significantly elevated by approximately 2.5- to 3.5-fold respectively. Similarly, D2B and D2K had some of the most highly expanded integrations in *CTDSPL2* and we observed a corresponding 4.5 fold increase in expression (Figure 5.3B). It is interesting to note that tumors in D2 have clonally expanded integrations in both genes but only one of the genes is overexpressed.



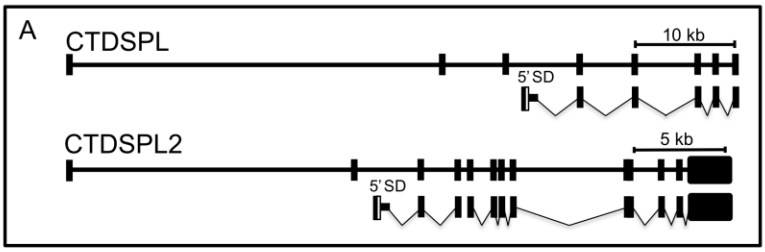

n = 3; **p < 0.0005, *p < 0.05

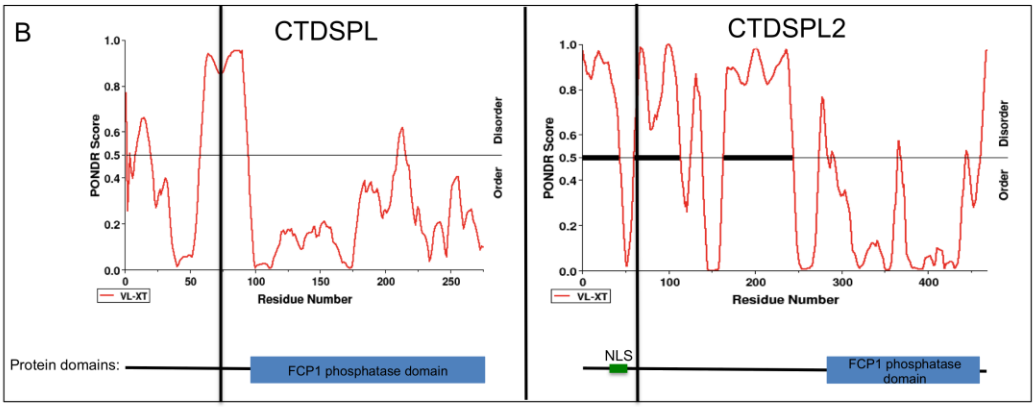
**Figure 5.3: Tumors with expanded integrations in *CTDSPL* and *CTDSPL2* overexpress transcripts**. (A) qPCR was performed from tumor cDNA for either *CTDSPL* or *CTDSPL2* and normalized to a housekeeping gene, GAPDH. Fold change in mRNA expression is depicted relative to expression levels in normal bursa (NB). Tumors with the most highly expanded integrations in *CTDSPL* significantly overexpress *CTDSPL* mRNA by 3- to 3.5- fold (C3L and C3K). (B) Likewise, those tumors with the most expanded integrations in *CTDSPL2* also have a 4.5-fold increase in *CTDSPL2* mRNA expression (D2B and D2K).

*Integrations in CTDSPL and CTDSPL2 generate truncated fusion transcripts.*

To determine the mechanism by which the viral integrations are disrupting *CTDSPL* and *CTDSPL2* expression, we performed RT-PCR to detect any potential viral fusion transcripts. We found that integrations in *CTDSPL* were driving the expression of a fusion transcript from the viral promoter with splicing occurring from the canonical splice donor site in *gag* to the splice acceptor site of exon 4 of the *CTDSPL* mRNA removing 77 amino acids from the N-terminus of the protein (Figure5. 4A). Integrations in *CTDSPL2* were driving expression of a fusion transcript from the viral promoter with splicing occurring from the canonical splice donor site in *gag* to the splice acceptor site of exon 3 of *CTDSPL2* removing 63 amino acids from the N-terminus of the protein (Figure 5.4A). In both cases, the viral start codon was in frame with the open reading frames and would add 6 amino acids of ALV *gag* at the N-terminus of the fusion protein. The truncation did not affect the catalytic phosphatase domain of either protein but did remove a portion of a predicted intrinsically disordered region of both proteins (Figure 5.4B). In the case of CTDPSL2, the truncation also removed a predicted nuclear localization signal (NLS; Figure 5.4B).

**Figure 5.4: *CTDSPL* and *CTDSPL2* transcript truncations induced by viral integrations**. (A) Schematic of truncated transcripts expressed from integrations in *CTDSPL* and *CTDSPL2* as detected by RT-PCR. The promoter in the 5' LTR of ALV drives expression of truncated transcript. Transcripts contain the *gag* leader sequence spliced from the canonical splice donor site (5' SD) into either exon 4 of *CTDSPL* or exon 3 of *CTDSPL2*. The ORF is not disrupted in either truncated transcript. (B) PONDR plots of CTDSPL and CTDSPL2. PONDR was used to predict intrinsically disordered regions of both CTDSPL and CTDSPL2 proteins (Xue et al., 2010). PONDR VL-XT score is indicated by red line. Threshold for disorder is set at 0.5 and indicated by a horizontal line. Significant stretches of disorder are indicated by thick black horizontal lines. The N-terminal portion of CTDSPL2 is significantly more disordered than CTDSPL. Truncations induced by viral integrations are indicated by a vertical black line. In both CTDSPL and CTDSPL2, the truncations remove a portion of the predicted disordered region. For CTDSPL2, the truncation also removes a predicted nuclear localization signal (NLS). For reference the catalytic domain (blue box) and NLS (green) are shown in a schematic representation of CTDSPL and CTDSPL2 at the bottom of PONDR plots.


A
CTDSPL
10 kb
5' SD
CTDSPL2
5 kb
5' SD


B
CTDSPL
CTDSPL2
PONDR Score
Disorder
Order
VL-XT
Residue Number
Protein domains:
FCP1 phosphatase domain
NLS
FCP1 phosphatase domain

*CTDSPL and CTDSPL2 induce expression changes in genes implicated in cellular migration, translation, alternative splicing and oxidative phosphorylation*

To better characterize the role of *CTDSPL* and *CTDSPL2* in ALV-induced B-cell lymphomas, we generated truncated transcripts in viral vectors to mimic those being expressed in tumors. Chick embryo fibroblasts (CEF) were infected with retroviral vectors (RCAS(A)) carrying either the truncated or full-length transcript of either *CTDSPL* or *CTDSPL2*. Transcripts were overexpressed approximately 100-fold relative to wild type CEF expression.

Both CTDSPL and CTDSPL2 are believed to act on the CTD of RNA polymerase II to regulate gene expression (Yeo et al., 2003). We reasoned that overexpression of these genes by viral integration may be affecting downstream gene expression. To identify changes in gene expression, RNA-seq analysis was performed on cells overexpressing truncated or full length *CTDSPL* or *CTDSPL2*. Cufflinks (Trapnell et al., 2012) was used to detect genes differentially expressed in cells carrying a *CTDSPL* or *CTDSPL2* construct relative to cells infected with an empty retroviral construct.

We observed between 4 and 30 genes differentially expressed in each condition (Figure 5.5, Table 5.3). There was very little overlap in differentially expressed genes between overexpression conditions. MMP9, or matrix metalloproteinase-9 is the only gene that was significantly deregulated by overexpression of all constructs. Cells expressing full length *CTDSPL* or *CTDSPL2* had the most similar changes in gene expression profiles with approximately 1/3 of the deregulated genes overlapping between the two conditions, suggesting that they may play partially redundant roles.

In order to determine differences in gene regulation induced by truncation of *CTDSPL* or *CTDSPL2* genes we performed a GO analysis of genes differentially expressed between the full length and truncated form of both *CTDSPL* and *CTDSPL2* separately (Table 5.2; Table 5.4). Differentially regulated genes between truncated and full length *CTDSPL* were enriched for genes involved in mitochondria, oxidative phosphorylation, alternative splicing and Sp1 targets. Mitochondrial genes as well as genes involved in oxidative phosphorylation were found to be upregulated in the cells expressing truncated *CTDSPL*. Sp1 target genes and genes involved in alternative splicing were found to be downregulated in cells expressing the truncated construct.

In both cases an enrichment of genes involved in cellular locomotion or focal adhesion, E2F targets and ribosomal genes were observed. There was no clear trend of upregulation or downregulation of the genes in these GO categories. There was little overlap in affected genes between full length and truncated constructs. Expression of truncated *CTDSPL* or *CTDSPL2* induced fewer changes in gene expression than either full-length construct indicating that the truncation may result in a partial loss of function.

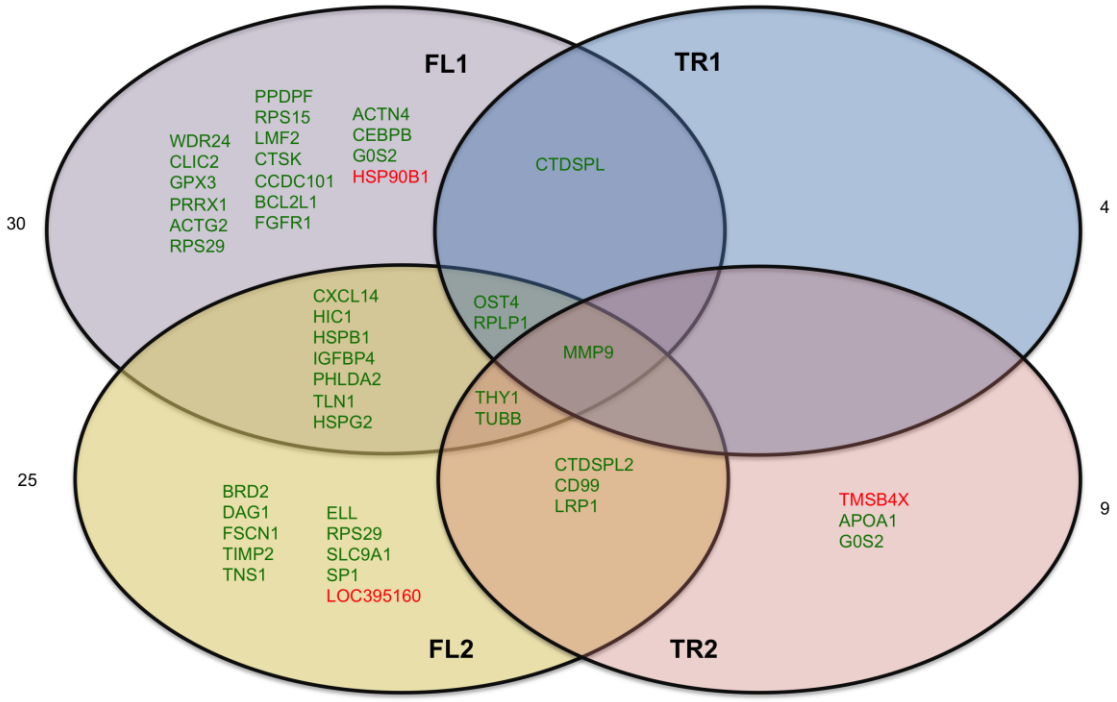



**Figure 5.5: Genes differentially expressed by overexpression of *CTDSPL* or *CTDSPL2* full length or truncated transcripts**. RNA-seq analysis of CEF cells overexpressing truncated or full length *CTDSPL* or *CTDSPL2* revealed a number of significantly overexpressed genes. Venn diagram depicting deregulation of gene expression by overexpression of truncated or full length CTDSPL or CTDSPL2. Genes shown in green were significantly upregulated relative to cells infected with an empty viral vector. Genes shown in red were significantly downregulated. Fold change in expression for each gene is given in Table 5.3.

| GO term | P-value |
|---|---|
| ***Enrichment in genes upregulated in TR1 vs. FL1*** | |
| Metabolism | 0.00000125 |
| Oxidative phosphorylation | 0.00000301 |
| Ribosome | 0.000927 |
| Adherens/anchoring junctions | 0.00471 |
| Factor: E2F | 0.00997 |
| ***Enrichment in genes downregulated in TR1 vs. FL1*** | |
| Focal adhesion | $6.47x10^{-10}$ |
| Factor: ETF | 0.00000002 |
| Factor: Sp1 | 0.00000026 |
| Factor: E2F-3 | 0.00000897 |
| Cell migration | 0.000286 |
| Alternative splicing | 0.0000654 |
| ***Enrichment in genes upregulated in TR2 vs. FL2*** | |
| Ribosome | 0.046 |
| Cell adhesion | 0.05 |
| ***Enrichment in genes downregulated in TR2 vs. FL2*** | |
| Focal adhesion | 0.0062 |
| Ribsomal protein | 0.0157 |

**Table 5.2: Gene ontology (GO) analysis of genes differentially regulated by overexpression of truncated versus full length CTDSPL or CTDSPL2.**

*CTDSPL or CTDSPL2 expression induces cell migration in vitro*

Due to the enrichment of genes involved in migration, such as *MMP9*, we next looked at whether cells overexpressing full length or truncated *CTDSPL* or *CTDSPL2* had any differences in ability to migrate. To do this, we made use of a wound healing assay, or scratch assay, in which a confluent plate of CEF cells was scratched to disrupt the monolayer. At subsequent times after inflicting the "wound", cells were imaged to visualize cell migration (Figure 5.6A). We observed that cells expressing either full length or a truncated *CTDSPL* or *CTDSPL2* transcript had a significantly higher rate of cell migration compared to an empty vector control.

Cells migrating into the wound were quantified, and percent wound closure was calculated (Figure 5.6B). *CTDSPL2* full-length overexpression had the largest effect with 25% wound closure compared to just 5% closure seen in the empty vector control ($p < 0.0001$). The truncated form of *CTDSPL2* had a more modest effect with 18% closure observed on average ($p < 0.0001$; Figure 5.6B). Cells expressing *CTDSPL* truncated and full-length transcripts had intermediate migration rates.

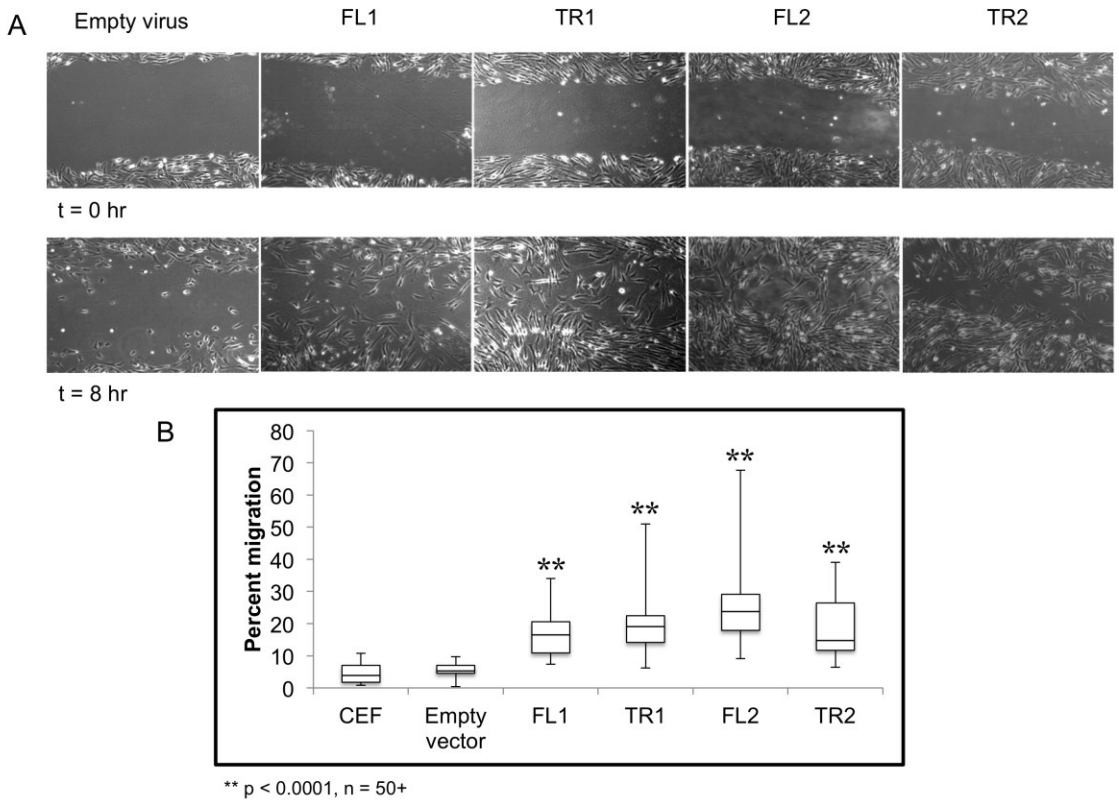



**Figure 5.6: *CTDSPL* and *CTDSPL2* promote cellular migration in chick embryo fibroblasts**. (A) A scratch assay was performed to monitor cell migration. Representative images of scratches at time 0 and at 8 hours are shown. (B) Quantification of wound closure. On average, uninfected CEF and CEF infected with empty viral vector exhibit approximately 5% wound closure after 8 hours. Cells expressing truncated *CTDSPL2* (TR2) exhibited significantly faster cellular migration rates with 18% wound closure on average after 8 hours ($p < 0.0001$). Cells expressing full length *CTDSPL2* (FL2) had the fastest migration with 25% wound closure at the final time point ($p < 0.0001$). Cells overexpressing CTDSPL truncated and full-length (TR1, FL1) transcripts had intermediate phenotypes with approximately 15-20% migration.

*CTDSPL2 overexpression prevents apoptosis induced by oxidative stress*

Promotion of cell migration by *CTDSPL* and *CTDSPL2* overexpression was observed when either truncated or full-length transcripts were expressed. Thus, this function does not explain why viral integrations that induce truncations were selected for in the tumors that we analyzed. Integrations in genes may also be selected for because they promote survival. To determine if integrations in *CTDSPL* and *CTDSPL2* are affecting survival, we induced apoptosis in cells expressing either full length or truncated *CTDSPL* or *CTDSPL2* by hydrogen peroxide treatment and measured cell death. Interestingly after 48 hours, cells expressing truncated or full length *CTDSPL2* had significantly higher survival rates than cells expressing *CTDSPL* or empty vector control. CEF cells expressing either form of *CTDSPL2* had approximately 3-fold higher survival than cells infected with an empty vector control (Figure 5.7).

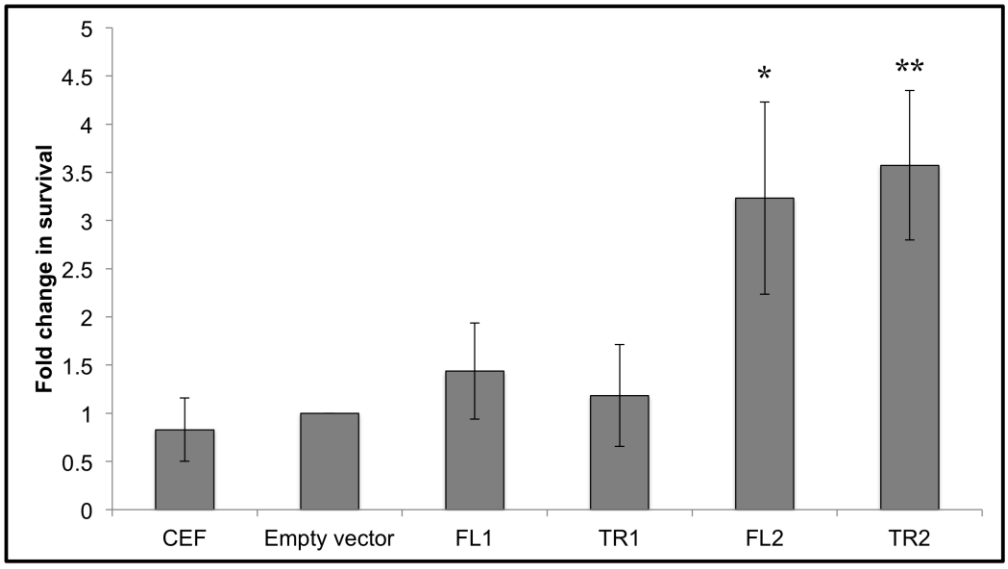



**Figure 5.7: *CTDSPL2* protects cells from apoptosis *in vitro*.** Chick embryo fibroblasts were treated with hydrogen peroxide to induce apoptosis and survival was measured relative to cells infected with an empty viral vector. Expression of either full-length or truncated *CTDSPL* (FL1, TR1) provided no protection from apoptosis ($p < 0.05$). Cells expressing either truncated or full length *CTDSPL2* (TR2, FL2) showed an approximately 3-fold increase in cell survival ($p< 0.05$).

*Overexpression of truncated viral fusion CTDSPL or CTDSPL2 transcripts promotes immortalization of primary cells in culture*

The typical lifespan of primary chicken embryo fibroblasts in culture is approximately 30 days. After this point, proliferation of CEF cells as well as ALV-infected CEF cells decreases dramatically. Overexpression of either full length *CTDSPL* or *CTDSPL2* did not affect proliferation at later time points. In contrast, cells overexpressing the viral fusion transcripts of *CTDSPL* and *CTDSPL2* did not undergo senescence (Figure 5.8). These cells continued proliferating at the same rate that was observed at earlier time points (data not shown). This effect of the truncated products on immortalization is likely the reason integrations were selected for in our initial screen of ALV-induced B-cell lymphomas.

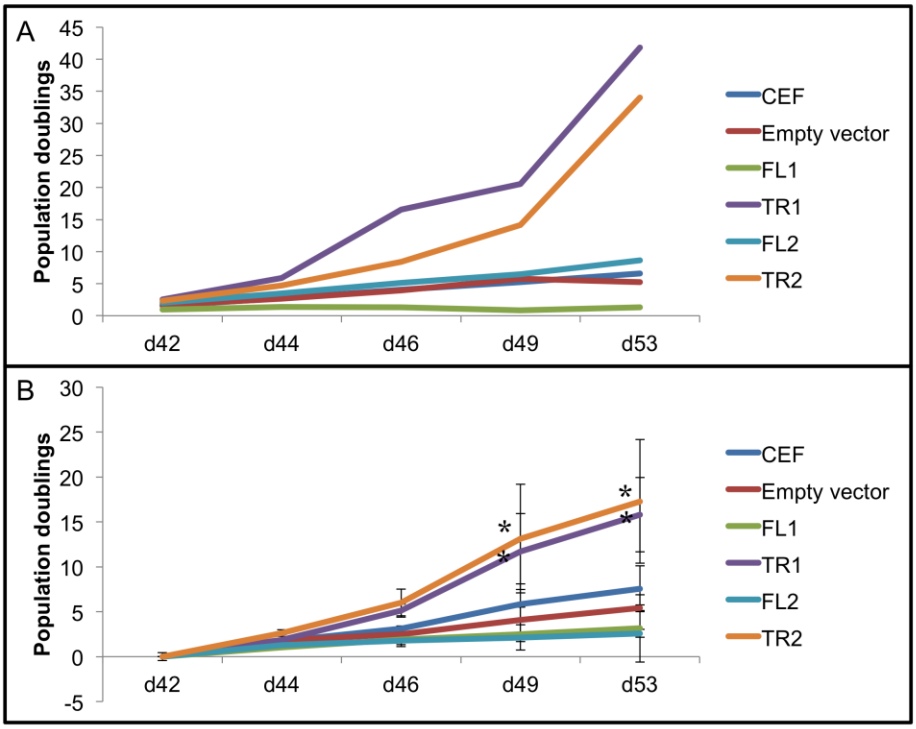



**Figure 5.8: Overexpression of truncated *CTDSPL* and *CTDSPL2* promotes immortalization of primary cells in culture**. Proliferation data shown is from day 42 to day 53 after infection. (A) Representative growth curve of CEF cells infected with virus, truncated or full length *CTDSPL* or *CTDSPL2*. (B) Average growth curve of three biological replicates. CEF cells as well as cells infected with empty virus stop proliferating at later time points. Cells overexpressing full length *CTDSPL* or *CTDSPL2* (FL1, FL2) proliferate less on average than uninfected CEFs. However, cells overexpressing truncated viral fusion transcripts of either *CTDSPL* or *CTDSPL2* (TR1, TR2) exhibit significantly higher proliferation at later time points indicating that they may be promoting immortalization.

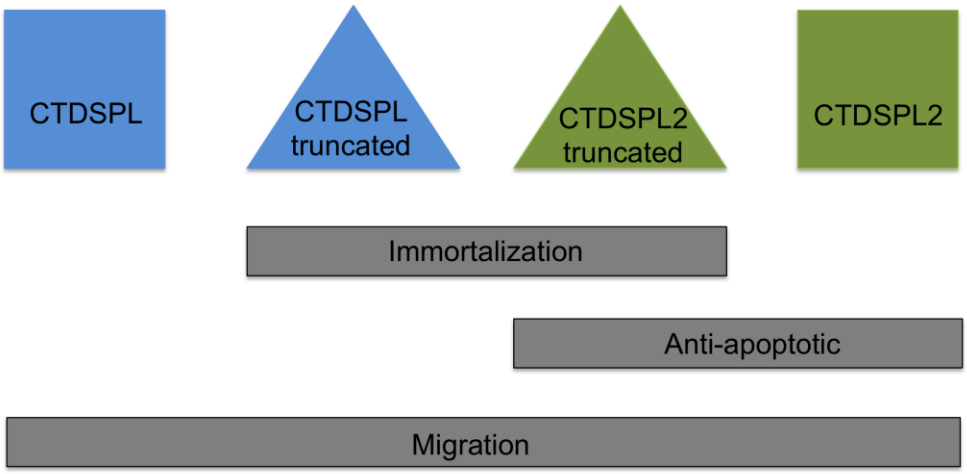



**Figure 5.9: Summary of findings**. Overexpression of the truncated viral fusion transcripts of *CTDSPL* and *CTDSPL2* promote immortalization in primary cell culture. Expression of either full length or truncated *CTDSPL2* protected cells from apoptosis. Overexpression of all constructs caused a significant increase in cellular migration.

## 5.3 Discussion

In this report we have identified both *CTDSPL* and *CTDSPL2* as common integration sites in ALV-induced B-cell lymphomas. In addition to being common integration sites, a large number of integrations in these genes were clonally expanded, suggesting a role in tumorigenesis. Further evidence for a driving role in cancer is suggested by the presence of identical integrations in primary and secondary tumors within the same bird. This indicates that these integrations were likely an early event in the development of cancer within these birds.

We show that viral integrations in *CTDSPL* and *CTDSPL2* were driving the overexpression of a truncated transcript. Overexpression of *CTDSPL* and *CTDSPL2* caused changes in the expression of genes involved in cellular migration, most notably *MMP9,* which was upregulated by overexpression of all constructs. Correspondingly, we observed an increase in cellular migration rates in cells overexpressing truncated and full-length transcripts. This, in addition to the observation that integrations in *CTDSPL* and *CTDSPL2* occur in both primary and secondary tumors, suggests a potential role in promoting tumor metastasis. While CTDSPL2 is not well studied, it has been demonstrated to play a role in bone morphogenetic protein (BMP) signaling through dephosphorylation of Smad proteins (Zhao et al., 2014). This has been shown to strongly promote cell migration in hepatocellular carcinoma cell lines (Maegdefrau and Bosserhoff, 2012). Further, inhibition of BMP signaling suppressed metastasis in mammary cancer (Owens et al., 2015). This role of CTDSPL and CTDSPL2 in cellular migration agrees with previously published data that CTDSP1/2/L proteins promote migration through the activation of the SNAIL1 protein, a key regulator of migration (Wu

et al., 2009). The promotion of cellular migration appears to be a gain of function due to overexpression of the *CTDSPL* and *CTDSPL2* transcripts.

The overexpression of truncated viral fusion transcripts of both *CTDSPL* and *CTDSPL2* promotes immortalization of primary cells in culture. This is a feature unique to the truncated transcripts, as overexpression of full-length forms of both genes did not significantly improve proliferation rates at times past the normal lifespan of CEFs. We believe that this role in immortalization is likely the reason that integrations promoting the expression of truncated forms of both genes are selected for in ALV-induced B-cell lymphomas. This role in immortalization for CTDSPL and CTDPSL2 is interesting to note due to the co-occurrence of CTDSPL and CTDSPL2 integrations with integrations into TERT, which has previously been reported to promote immortalization (Bodnar et al., 1998).

CTDSP1/2/L are fairly well characterized genes that have been repeatedly shown to play partially redundant roles. CTDSPL2 seems to be fairly similar to the other members of the CTDSP family in many regards. Some functions are known to overlap, such as regulation of BMP signaling. Here we show that CTDSPL2 promotes metastasis similar to CTDSPL. These overlapping functions, in addition to the observation that despite integrations in both genes, only one is overexpressed in individual tumors, would suggest that CTDSPL and CTDSPL2 might be redundant. However, we observed that expression of *CTDSPL2*, and not *CTDSPL*, can protect cells from apoptosis induced by oxidative stress.

CTDSPL2 does have distinct features from the other members of the CTDSP family. For instance, CTDSP1/2/L genes have an intronic microRNA from the miR-26

family. No intronic microRNA has been reported in *CTDSPL2*. The CTDSPL2 protein, at 53 kDa, is significantly larger in size than CTDSP1/2/L proteins, which all weigh in at around 32 kDa on average. Each protein in the family contains a C-terminal phosphatase domain, but CTDSPL2 has significantly more N-terminal sequence of unknown function. The N-terminal region that is truncated in the viral fusion transcript is predicted to be intrinsically disordered (Figure 5.4B). Likewise, the truncated portion of CTDSPL is also predicted to be partially disordered. Given that the CTD of RNA polymerase II has been shown to associate with proteins with low-complexity domains, it is possible that these disordered regions are in part responsible for binding of CTDSPL and CTDSPL2 to the CTD. Thus, in the truncated transcript that is expressed in tumors, these proteins may not be able to associate with RNA polymerase II or other substrates.

Furthermore, unlike CTDSP1/2/L proteins, CTDSPL2 has not been previously reported to act on pRb, a main tumor suppressor target common to the other 3 members of the family. However, in our RNA-seq analysis, nearly 75% of the genes that were differentially expressed by 2-fold or more in cells overexpressing *CTDSPL2* were E2F target genes. E2F is a transcription factor targeted by pRb. When pRb is dephosphorylated and thus active, it binds E2F, keeping it inactive. Once pRb becomes phosphorylated in G1, it releases E2F allowing it to act on downstream effector genes causing the transition from G1 to S phase (Polager and Ginsberg, 2008). CTDSP1/2/L were shown to dephosphorylate and thus activate pRb (Zhu et al., 2012b) . Due to regulation of E2F target genes by CTDSPL2, it seems likely that this protein may also act as a phosphatase on pRb.

Our work suggests that *CTDSPL* and *CTDSPL2* play a role in cancer and seem to have pro-oncogenic characteristics (Figure 5.9). Expression of either of these genes promotes metastasis in cell culture and CTDSPL2 protects cells from apoptosis. Neither of these functions is affected by the viral truncation. We believe that the main reason integrations in *CTDSPL* and *CTDPSL2* were selected for in B-cell lymphomas is due to the role of the truncated transcripts in immortalization. We hypothesize that the gene truncations imposed by the viral integrations in tumors remove a region of the protein that is responsible for interaction with pRb. The truncated proteins would no longer be able to dephosphorylate pRb and would potentially lose their tumor suppressor function. Genes deregulated by expression of the truncated transcript were also enriched in downstream effectors and processes of the pRb pathway, such as E2F and Sp1 target genes (Polager and Ginsberg, 2008). This suggests that the truncated version of the proteins interact with pRb differently causing a change in expression of downstream effectors of pRb. We hypothesize that the removal of a portion of a predicted intrinsically disordered region may inhibit these proteins from interacting with its normal protein-binding partners. For CTDSPL2, the truncation also removes a nuclear localization signal that may prevent the protein from reaching the nucleus. pRb has been shown to be a dominant effector of cellular senescence with inactivation of pRb delaying onset of cellular senescence (Campisi, 2005; Haferkamp et al., 2009). If the truncated CTDSPL and CTDSPL2 proteins can no longer activate pRb through dephosphorylation, then pRb may become phosphorylated and thus inactive, allowing for evasion of senescence as observed in our cell culture system.

**Supplementary tables:**

| gene | sample_1 | sample_2 | FPKM_value_1 | FPKM_value_2 | log2(fold_change) | q_value |
|---|---|---|---|---|---|---|
| CTDSPL | virus | fl1 | 76.78 | 17615.00 | 7.84 | 0.01 |
| IGFBP4 | virus | fl1 | 97.05 | 344.36 | 1.83 | 0.01 |
| PHLDA2 | virus | fl1 | 848.26 | 2464.20 | 1.54 | 0.01 |
| WDR24 | virus | fl1 | 31.34 | 117.73 | 1.91 | 0.01 |
| HSPB1 | virus | fl1 | 453.66 | 1323.48 | 1.54 | 0.01 |
| CLIC2 | virus | fl1 | 77.73 | 245.47 | 1.66 | 0.01 |
| RPLP1 | virus | fl1 | 1633.35 | 4518.34 | 1.47 | 0.01 |
| MMP9 | virus | fl1 | 66.67 | 203.62 | 1.61 | 0.01 |
| GPX3 | virus | fl1 | 230.60 | 638.03 | 1.47 | 0.01 |
| PRRX1 | virus | fl1 | 402.33 | 983.46 | 1.29 | 0.01 |
| ACTG2 | virus | fl1 | 602.16 | 1497.56 | 1.31 | 0.01 |
| RPS29 | virus | fl1 | 1852.45 | 4731.70 | 1.35 | 0.01 |
| THY1 | virus | fl1 | 1222.43 | 3083.13 | 1.33 | 0.01 |
| TLN1 | virus | fl1 | 62.13 | 148.89 | 1.26 | 0.01 |
| PPDPF | virus | fl1 | 337.37 | 882.52 | 1.39 | 0.01 |
| HSPG2 | virus | fl1 | 32.95 | 78.14 | 1.25 | 0.01 |
| RPS15 | virus | fl1 | 2775.50 | 6422.73 | 1.21 | 0.01 |
| LMF2 | virus | fl1 | 216.30 | 502.87 | 1.22 | 0.01 |
| CTSK | virus | fl1 | 175.35 | 443.76 | 1.34 | 0.02 |
| CCDC101 | virus | fl1 | 141.08 | 337.39 | 1.26 | 0.02 |
| BCL2L1 | virus | fl1 | 166.00 | 436.28 | 1.39 | 0.02 |
| OST4 | virus | fl1 | 2030.88 | 5113.26 | 1.33 | 0.02 |
| TUBB | virus | fl1 | 450.43 | 966.51 | 1.10 | 0.02 |
| FGFR1 | virus | fl1 | 60.02 | 141.34 | 1.24 | 0.03 |
| HSP90B1 | virus | fl1 | 494.94 | 239.69 | -1.05 | 0.04 |
| HIC1 | virus | fl1 | 38.04 | 101.29 | 1.41 | 0.04 |
| ACTN4 | virus | fl1 | 86.34 | 205.06 | 1.25 | 0.04 |
| G0S2 | virus | fl1 | 275.39 | 691.71 | 1.33 | 0.04 |
| CXCL14 | virus | fl1 | 302.04 | 635.73 | 1.07 | 0.04 |
| CEBPB | virus | fl1 | 63.24 | 192.42 | 1.61 | 0.04 |
| CTDSPL | virus | tr1 | 79.02 | 7386.97 | 6.55 | 0.04 |
| OST4 | virus | tr1 | 2089.22 | 6399.39 | 1.61 | 0.04 |
| RPLP1 | virus | tr1 | 1680.51 | 5375.79 | 1.68 | 0.04 |
| MMP9 | virus | tr1 | 82.64 | 155.26 | 0.91 | 0.04 |
| BRD2 | virus | fl2 | 96.89 | 244.08 | 1.33 | 0.01 |
| CD99 | virus | fl2 | 264.15 | 641.17 | 1.28 | 0.01 |
| CXCL14 | virus | fl2 | 316.59 | 634.03 | 1.00 | 0.01 |
| DAG1 | virus | fl2 | 72.18 | 180.77 | 1.32 | 0.01 |
| FSCN1 | virus | fl2 | 171.70 | 416.65 | 1.28 | 0.01 |

| | | | | | | |
|---|---|---|---|---|---|---|
| HIC1 | virus | fl2 | 39.87 | 116.75 | 1.55 | 0.01 |
| HSPB1 | virus | fl2 | 475.54 | 1263.80 | 1.41 | 0.01 |
| HSPG2 | virus | fl2 | 34.53 | 84.45 | 1.29 | 0.01 |
| IGFBP4 | virus | fl2 | 101.73 | 373.23 | 1.88 | 0.01 |
| MMP9 | virus | fl2 | 69.88 | 210.86 | 1.59 | 0.01 |
| OST4 | virus | fl2 | 2128.86 | 5479.19 | 1.36 | 0.01 |
| PHLDA2 | virus | fl2 | 889.13 | 1754.00 | 0.98 | 0.01 |
| RPLP1 | virus | fl2 | 1712.10 | 3737.50 | 1.13 | 0.01 |
| TLN1 | virus | fl2 | 65.12 | 134.75 | 1.05 | 0.01 |
| TUBB | virus | fl2 | 472.14 | 1302.85 | 1.46 | 0.01 |
| CTDSPL2 | virus | fl2 | 60.90 | 5416.11 | 6.47 | 0.01 |
| LRP1 | virus | fl2 | 113.33 | 214.77 | 0.92 | 0.01 |
| TIMP2 | virus | fl2 | 1532.28 | 3064.68 | 1.00 | 0.01 |
| THY1 | virus | fl2 | 1281.35 | 2382.98 | 0.90 | 0.02 |
| TNS1 | virus | fl2 | 58.86 | 118.23 | 1.01 | 0.02 |
| ELL | virus | fl2 | 18.57 | 56.76 | 1.61 | 0.03 |
| RPS29 | virus | fl2 | 1941.83 | 3992.96 | 1.04 | 0.03 |
| SLC9A1 | virus | fl2 | 51.29 | 128.46 | 1.32 | 0.04 |
| LOC395160 | virus | fl2 | 7591.35 | 4226.02 | -0.85 | 0.04 |
| SP1 | virus | fl2 | 29.98 | 95.87 | 1.68 | 0.04 |
| CD99 | virus | tr2 | 257.15 | 894.19 | 1.80 | 0.02 |
| CTDSPL2 | virus | tr2 | 59.27 | 6472.04 | 6.77 | 0.02 |
| MMP9 | virus | tr2 | 68.05 | 204.13 | 1.58 | 0.02 |
| THY1 | virus | tr2 | 1247.71 | 2585.62 | 1.05 | 0.02 |
| TMSB4X | virus | tr2 | 22900.10 | 10839.20 | -1.08 | 0.02 |
| TUBB | virus | tr2 | 459.79 | 1019.54 | 1.15 | 0.02 |
| G0S2 | virus | tr2 | 281.09 | 908.97 | 1.69 | 0.02 |
| LRP1 | virus | tr2 | 110.33 | 242.59 | 1.14 | 0.02 |
| APOA1 | virus | tr2 | 170.46 | 479.51 | 1.49 | 0.03 |

**Table 5.3: Cuffdiff results comparing gene expression in cells expressing either truncated or full length CTDSPL or CTDSPL2 relative to cells infected with an empty viral vector**. Each sample is compared relative to cells infected with an empty virus. FPKM values for each sample are listed as well as the $\log_2$(fold change) between the samples.

| GO term | P-value | Genes |
|---|---|---|
| ***Enrichment in genes upregulated in TR1 vs. FL1*** | | |
| **Metabolism** | 0.00000125 | PSMB1, MGST1, MSMO1, OAT, ASNS, RPL31, NUP37, NDUFB2, PSMA6, SQLE, PPP2CB, PRKAG2, RPL34, NDUFA8, RPL5, AMD1 ,UQCR11, PAICS, DCTD, COX4I1, RAN, NDUFB5, HADH, LBR, RPL37, GSTO1, ATP5G3, CMPK1, HPGDS, RPL22L1, STARD4, CYB5A, IDH3A, ACAA2, UQCRFS1, SLC25A6, BPGM, RPL38, UQCRH, FAR1, RPL12, SDHD |
| **Oxidative phosphorylation** | 0.00000301 | NDUFB2, NDUFA8, UQCR11, ATP6V1E1, COX4I1, NDUFB5, ATP6V1G1, ATP5G3, UQCRFS1, UQCRH, SDHD |
| **Ribosome** | 0.000927 | MRPS35, RPL31, RPL34, MRPS14, RPL5, RPL37, RPL22L1, RPL38, RPL12, MRPS17 |
| **Adherens/anchoring junctions** | 0.00471 | RPL31, CHMP2B, RPL34, PRDX1, RPL5, PAICS, RAN, RRAS2, ACTR2, ITGAV, G3BP1, TWF1, CCT8, ANXA5, RPL38, RPL12 |
| **Factor: E2F** | 0.00997 | PSMB1, GLT8D1, MSMO1, MRPS35, OAT, RPL31, WDR1, SLC25A3, PREP, NDUFB2, TXN2, PSMA6, PGRMC1, RP2, SQLE, PPP2CB, PRKAG2, UFSP2, TMEM30A, USP4, PRDX1, STMN1, UCHL3, NDUFA8, UFM1, TBC1D15, SPRYD7, AMD1, UQCR11, PAICS, DCTD, LSM7, ATP6V1E1, COX4I1, CHMP1A, RAN, TPT1, SLC38A2, CDC73, MPHOSPH6, ATP6V1G1, ACTR2, ITGAV, HADH, NEDD1, TSPAN3, CDR2, LBR, RPL37, G3BP1, GSTO1, N6AMT2, TWF1, UCHL1, VMA21, CMPK1, MEMO1, HPGDS, ATP1A1, SLMAP, H2AFZ, ANXA5, STARD4, VDAC2, CYB5A, IDH3A, PHB, SNRPD1, CRK, ACAA2, UQCRFS1, SLC25A6, ARL6IP1, RNF139, RAB33B, BPGM, RPL38, UQCRH, AP3S1, MCFD2, ALDH1A3, YTHDF3, GMFB, FAR1, CAPZA2, TMSB4X, VDAC1, MRPS17 |
| ***Enrichment in genes downregulated in TR1 vs. FL1*** | | |
| **Focal adhesion** | $6.47 \times 10^{-10}$ | CD99, ITGA3, PPP1R12A, PABPC1, TNS1, PPFIBP1, PTK7, LRP1, FLOT2, RAC1, TLN1, GIT2, HSPG2, THY1, ZYX, RPL8, ARF6, ADAM9, YES1, EVL, TGM2, MARCKS |
| **Factor: ETF** | 0.00000002 | CD99, ITGA3, LAMP2, MAP2K3, TIMP2, VCAN, FAM214A, MRTO4, PPP1R12A, HDAC7, WAPAL, NAV3, MEF2A, PABPC1, NDE1, PTGS2, FGFR1, SMARCA2, TDRD3, AK6, CCNK, FKBP5, KDELR3, CHMP4B, NDFIP2, TSC22D1, FAM173A, SF3A2, MEOX2, RSRC2, ELK3, CPSF6, RWDD1, PTK7, IK, CISH, NCL, SPTBN1, HPCAL1, PRRX1, B4GALT2, PRDX6, EGR1, CLU, LRP1, G0S2, STAU1, PPDPF, WDR24, WDR44, FLOT2, MYH10, SRRM1, BHLHE40, |

| | | |
|---|---|---|
| | | PUM1, FADS2, RAC1, TLN1, ETV6, GIT2, MFAP1, PARN, IGFBP4, COL6A1, COL6A2, HSPG2, RIT1, SNRNP200, METTL21A, CTDSPL, ADPRH, CCNA2, SFRP2, CXCL14, ADD3, ADAM33, HNRNPDL, CFDP1, THY1, ELMO1, EIF5B, RNF166, ZYX, RPL8, PPAP2B, ZNF326, PM20D1, RPRD2, CDCA7L, ARF6, TAF3, MAPRE2, LINGO1, SIX2, PDGFD, CANT1, HDAC3, PTPN2, YES1, BASP1, PTRF, PHLDA2, PRKAG1, COL18A1, EWSR1, TOP1MT, PDE4B, SP1, COL4A1, NOC2L, H1F0, SRC, SPG7, MAFK, BRD2, GPX3, MARCKS |
| **Factor: Sp1** | 0.00000026 | CD99, ITGA3, LAMP2, MAP2K3, TIMP2, VCAN, FAM214A, PPP1R12A, HDAC7, NAV3, MEF2A, PABPC1, NDE1, FGFR1, TNS1, SMARCA2, AK6, CCNK, FKBP5, KDELR3, CHMP4B, NDFIP2, TSC22D1, FAM173A, SF3A2, CBLL1, PPFIBP1, RSRC2, ELK3, CPSF6, PTK7, IK, CISH, NCL, SPTBN1, HPCAL1, PRRX1, B4GALT2, PRDX6, DUSP1, EGR1, CLU, CDK2, LRP1, STAU1, PPDPF, WDR44, SERPINF1, FLOT2, NES, MYH10, SRRM1, BHLHE40, FADS2, RAC1, TLN1, ETV6, MFAP1, MFGE8, PARN, IGFBP4, COL6A1, COL6A2, HSPG2, RGL1, RIT1, SNRNP200, METTL21A, CTDSPL, ADPRH, CCNA2, SFRP2, CXCL14, ADD3, ADAM33, HNRNPDL, CFDP1, THY1, EIF5B, RNF166, RUNX1, ZYX, RPL8, PPAP2B, ZNF326, PM20D1, RPRD2, ARF6, MAPRE2, ADAM9, UGP2, LINGO1, SIX2, PDGFD, CANT1, BCL2L1, HDAC3, PTPN2, YES1, CCDC101, BASP1, PTRF, PHLDA2, PRKAG1, COL18A1, EWSR1, TOP1MT, PDE4B, SP1, COL4A1, NOC2L, H1F0, SUPT5H, SRC, SPG7, MAFK, TGM2, BRD2, GPX3, MARCKS |
| **Factor: E2F-3** | 0.00000897 | CD99, ITGA3, LAMP2, MAP2K3, TIMP2, VCAN, FAM214A, HDAC7, WAPAL, NAV3, MEF2A, PABPC1, NDE1, FGFR1, SMARCA2, TDRD3, AK6, CCNK, FKBP5, KDELR3, CHMP4B, TSC22D1, FAM173A, SF3A2, PPFIBP1, ELK3, CPSF6, PTK7, IK, CISH, NCL, SPTBN1, HPCAL1, PRRX1, B4GALT2, PRDX6, DUSP1, EGR1, CLU, CDK2, LRP1, STAU1, PPDPF, WDR44, SERPINF1, FLOT2, NES, MYH10, SRRM1, BHLHE40, FADS2, RAC1, TLN1, ETV6, MFAP1, MFGE8, PARN, IGFBP4, COL6A1, COL6A2, HSPG2, RGL1, RIT1, METTL21A, CCNA2, ADD3, ADAM33, HNRNPDL, CFDP1, THY1, ELMO1, RNF166, RUNX1, ZYX, RPL8, ZNF326, PM20D1, RPRD2, CDCA7L, ARF6, TAF3, MAPRE2, ADAM9, UGP2, LINGO1, SIX2, CANT1, BCL2L1, HDAC3, PTPN2, YES1, |

| | | |
|---|---|---|
| | | CCDC101, BASP1, PTRF, NQO1, PHLDA2, PRKAG1, COL18A1, EWSR1, TOP1MT, SP1, COL4A1, NOC2L, H1F0, SRC, SPG7, MAFK, TGM2, BRD2, GPX3, MARCKS |
| **Cell migration** | 0.000286 | ITGA3, HDAC7, NDE1, PTGS2, FGFR1, TNS1, CBLL1, MEOX2, ARHGDIB, PTK7, LRP1, SERPINF1, MYH10, RAC1, SFRP2, CXCL14, THY1, ELMO1, PPAP2B, ADAM9, SIX2, PDGFD, YES1, PHLDA2, COL18A1, PDE4B, EVL, SRC |
| **Alternative splicing** | 0.0000654 | TSC22D1, LAMP2, LRP1, CTDSPL, COL6A2, PPP1R12A, PUM1, VCAN, BCL2L1, SRC |
| ***Enrichment in genes upregulated in TR2 vs. FL2*** | | |
| **Ribosome** | 0.046 | RPS2, MRPL33, RPS3 |
| **Cell adhesion** | 0.05 | RPS2, MRPL33, RPS3 |
| ***Enrichment in genes downregulated in TR2 vs. FL2*** | | |
| **Focal adhesion** | 0.0062 | CAV1, RPS29, RPLP1, DAG1, CAPN2 |
| **Ribsomal protein** | 0.0157 | MRPS26, RPLP1, MRPL20 |

**Table 5.4: Detailed gene ontology (GO) information**. GO terms that were enriched for genes up- or downregulated in cells overexpressing the truncated form of the gene versus full length. Genes that fall into each GO category are listed.

# Chapter 6 – ALV activation of a novel antisense RNA upstream of *TERT* in B-cell lymphomas

Adapted from:
Nehyba J.*, Malhotra S.*, **Winans S**.*, O'Hare T, Justice J. 4th, Beemon K. (2016). Avian leukosis virus activation of an antisense RNA upstream of *TERT* in B-cell lymphomas. *J Virol*. 90(20):9509-17.
(* indicates co-first authors).

**Summary**

Avian leukosis virus (ALV) induces tumors by integrating its proviral DNA into the chicken genome and altering expression of nearby genes via strong promoter and enhancer elements. Viral integration sites that contribute to oncogenesis are selected in tumor cells. Deep sequencing analysis of B-cell lymphoma DNA confirmed that the telomerase reverse transcriptase (*TERT*) promoter is a common ALV integration target. Twenty-six unique proviral integration sites were mapped between 46 and 3552 nt upstream of the *TERT* transcription start site, predominantly in the opposite transcriptional orientation of *TERT*. RNA-seq analysis of normal bursa revealed a transcribed region upstream of *TERT* in the opposite orientation, suggesting the *TERT* promoter is bidirectional. This transcript appears to be an uncharacterized antisense RNA which we have named *TAPAS* (TERT antisense promoter associated) *RNA*. We have previously shown that *TERT* expression is up regulated in tumors with integrations in the *TERT* promoter region. We now report that the viral promoter drives expression of a chimeric transcript, containing viral sequences spliced to exons 4 through 7 of this antisense RNA. Clonal expansion of cells with ALV integrations driving over expression of this *TERT* antisense RNA suggest it may have a role in tumorigenesis. Functional analysis of the *TAPAS RNA* transcript reveal that it plays a role in regulating TERT mRNA expression in chickens. We also find evidence of a similar transcript in humans that likewise appears to play a role in regulating TERT expression.

## 6.1 Introduction

High throughput sequencing revealed multiple integration sites in a series of rapid-onset ALV-induced B-cell lymphomas (6). The *TERT* promoter region was identified as the most clonally expanded of these integrations, suggesting this is an early event in tumorigenesis (Justice et al., 2015b). Twenty-six unique integration sites were identified in this region in multiple independent tumors (Justice et al., 2015b).

Telomerase is a ribonucleoprotein complex that adds repeat sequences to chromosome ends. It contains a catalytic protein component, TERT, as well as a non-coding telomerase RNA template component (TERC). Elevated telomerase activity has been detected in more than 90% of all human cancers (Shay and Wright, 2011). In addition, many human tumors have a point mutation in the *TERT* promoter at -124 or -146 nt upstream of the *TERT* translation start site (Heidenreich et al., 2014). These mutations up-regulate *TERT* expression (Borah et al., 2015; Horn et al., 2013; Huang et al., 2013; Killela et al., 2013). Elevated telomerase activity maintains telomere lengths and prevents apoptotic signaling, thus allowing continual proliferation and long-term viability of cancer cells (Blasco, 2005). It has also been shown that *TERT* can promote oncogenesis independent of the reverse transcriptase function of telomerase (Koh et al., 2015).

Telomerase activity in most somatic cells is limited by the availability of TERT protein, and expression of *TERT* is tightly regulated at the transcriptional level through epigenetic modifications in the promoter region (Delany and Daniels, 2004; Zhu et al., 2010). In addition, extensive alternative splicing of the *TERT* transcript has been shown to generate inactive variants that decrease telomerase activity (Hrdlicková et al., 2012;

Saebøe-Larssen et al., 2006; Withers et al., 2012). Both human and chicken telomerase expression is down regulated in most normal somatic tissues (Collins and Mitchell, 2002; Taylor and Delany, 2000). Furthermore, human and chicken telomeres shorten with age, and telomerase activity is important for oncogenesis (Delany et al., 2000). In contrast, mice express telomerase in normal somatic cells and have longer telomeres than humans or chickens (Hackett and Greider, 2002). Therefore, chicken serves as a good model to study oncogenic events of TERT activation and signaling.

We previously reported that ALV integrations upstream of *TERT* cause a 2-4 fold up-regulation of *TERT* expression in rapid-onset B-cell lymphomas (Yang et al., 2007b). However, these integrations were in the opposite transcriptional orientation to *TERT*, unlike most previously characterized common integration sites in ALV-induced tumors (Clurman and Hayward, 1989; Hayward et al., 1981; Kanter et al., 1988). In this work, we show that these integrations also drive over-expression of a novel antisense transcript, associated with the bidirectional *TERT* promoter, which we call *TAPAS* (TERT Antisense Promoter ASsociated) RNA. The ALV integrations result in chimeric transcripts with ALV leader sequences spliced into exon 4 of the 7-exon TAPAS RNA. Knockdown of this transcript causes a concordant decrease in *TERT* mRNA expression in primary chick embryo fibroblasts. Overexpression of TAPAS had little effect on TERT expression indicating that TAPAS is likely regulating *TERT* in *cis*.

Analysis of RNA-seq data revealed the presence of a transcript in humans in the same region of the *TERT* promoter. Using RT-PCR, we characterized a 1.6 kb transcript 170 nt upstream of the *TERT* transcription start site. It is an unspliced transcript and also contains no significant open reading frames. Contrary to that seen in CEF cells

knockdown of this hTAPAS transcript in HEK293T cells causes an increase in TERT mRNA expression. In agreement, analysis of cancer transcriptome data from various human patient samples from TCGA reveals that *hTERT* and *hTAPAS* expression are inversely correlated.

## 6.2 Results

*TERT promoter is a common ALV integration site in B-cell lymphomas*

In order to identify genes contributing to the formation of ALV-A induced rapid-onset B-cell lymphomas, high throughput sequencing of proviral – host DNA junctions was previously performed (Justice et al., 2015b). Common integration sites in the host genome that contribute to tumorigenesis are present in multiple tumor cells and thus are overrepresented in the deep sequencing data. The *TERT* promoter region was identified as the most clonally-expanded, common integration site with integrations present in seven different lymphomas from five birds (Figure 6.1A). We analyzed 19 of the most clonally expanded, unique integrations from both primary bursal tumors (B) and tumors metastasized to the liver (L). Three of the clonally expanded integrations were present in multiple tissues from the same bird. The integration sites ranged from 46 nt to 3552 nt upstream of the *TERT* transcription start site. The majority of the proviral integrations (16/19) were in the opposite transcriptional orientation to *TERT*. Four out of 7 lymphomas, termed C7B, C6L, C7L, and D2L had integrations only in the opposite orientation. The remaining three tumors – A1B, C2B, and C2L – harbored integrations in the same orientation as *TERT*, but also contained integrations in the opposite orientation that were more clonally expanded. In contrast, no integrations in the *TERT* promoter region were identified in any non-tumor tissues of infected birds (Figure 6.1B). The observation of proviral integrations in this region in multiple tumors suggests that ALV integration in the *TERT* promoter contributes to driving lymphomagenesis in these birds.

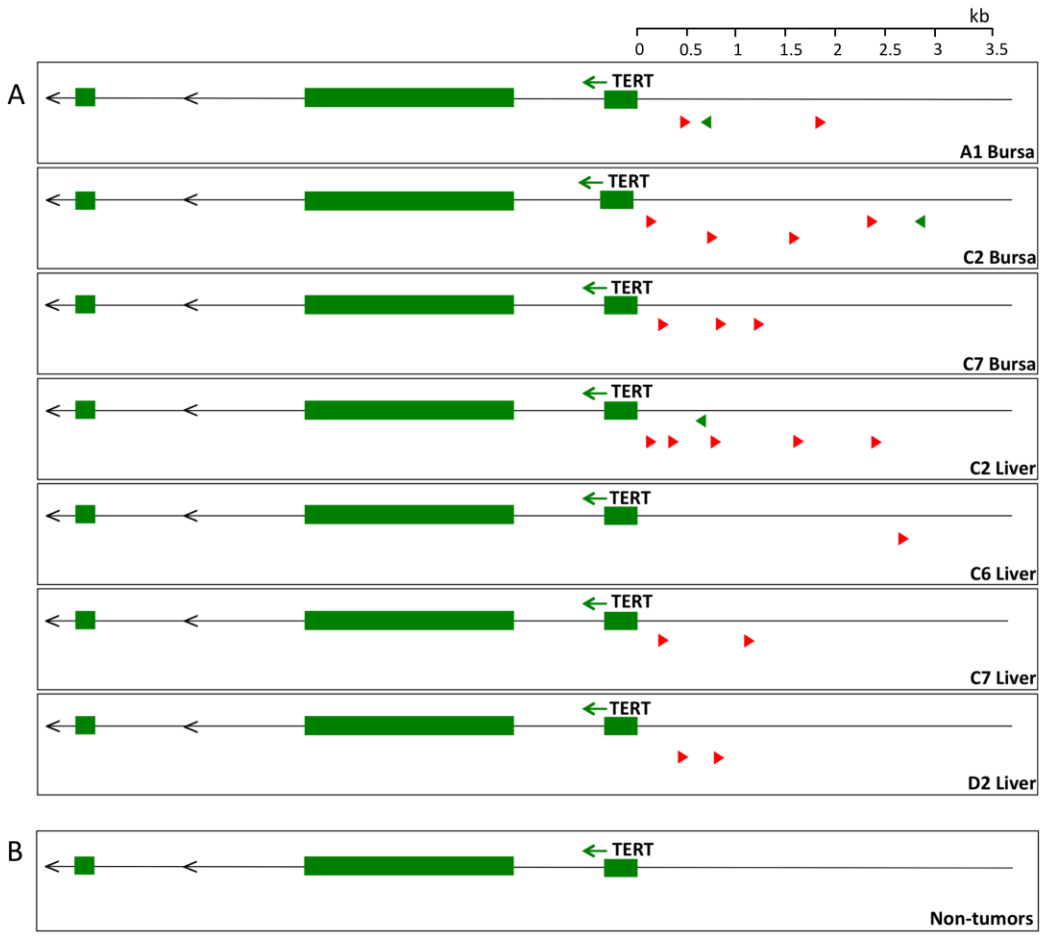



**Figure 6.1: The *TERT* promoter region is a common site of ALV proviral integration in lymphomas.** (A) Schematic of the most clonally expanded ALV integration sites near *TERT* in 7 tumors, shown with the first 3 exons of the *TERT* gene. Tumor names correspond to the bird and tissue in which the tumor was collected. All of the integrations clustered within 3.5 kb upstream of the *TERT* transcriptional start site. Integrations are predominantly in the opposite orientation (red) with respect to *TERT* transcription. (B) Schematic of integrations near *TERT* in 6 non-tumor tissues from infected chickens.

*Novel antisense (TAPAS) RNA is transcribed from TERT bidirectional promoter*

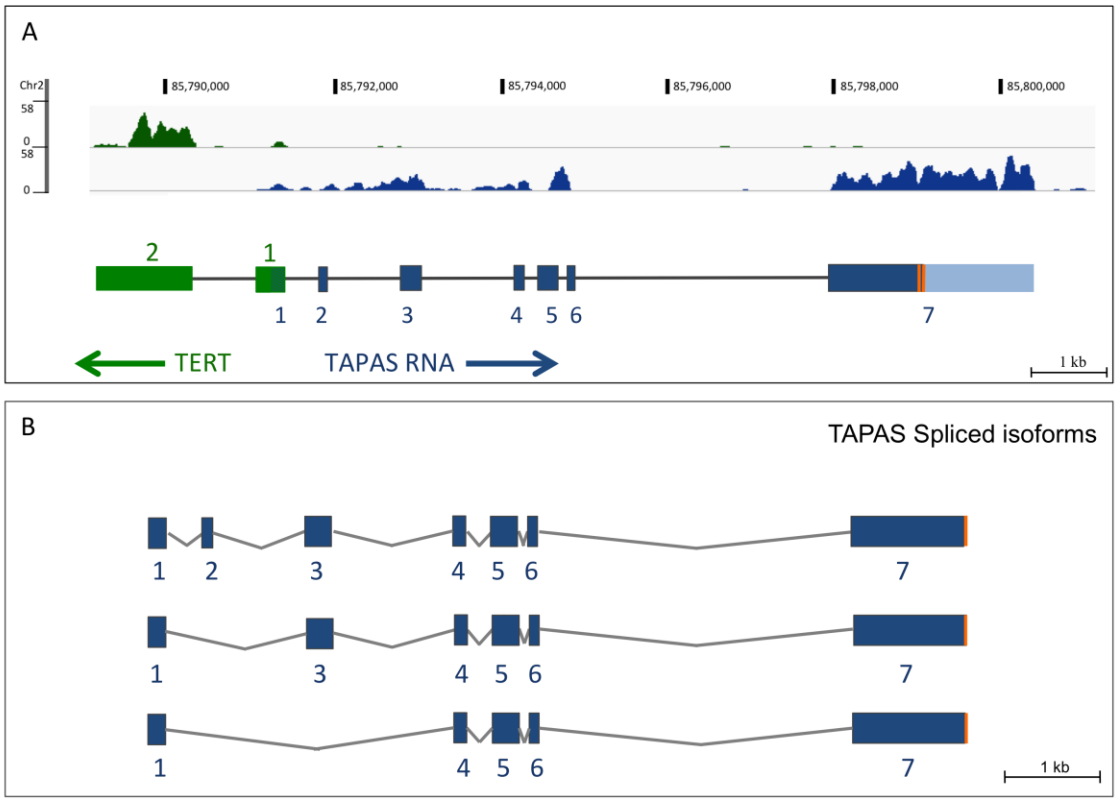
In order to assess the effects of proviral integrations on host gene expression, deep sequencing of the transcriptome of selected ALV-induced lymphomas and normal bursa controls was performed. This analysis revealed a 9 kb transcribed region upstream and in the opposite transcriptional orientation of *TERT* in the normal bursa controls (Figure 6.2A). This suggests that the *TERT* promoter is bidirectional. With the use of TopHat bioinformatics tools, a number of putative introns were identified and confirmed by sequence analysis of exon junctions. This analysis suggested a 3.6 kb spliced transcript, containing 7 exons. RT-PCR studies confirmed 2.2 kb of this transcript containing exons 1 through 7 (Figure 6.2A). RT-PCR experiments were not able to amplify the last 1050 nt of exon 7. Two putative poly(A) sites were identified by 3'RACE at positions 1051 and 1114 of exon 7.

Strand-specific RNA-seq data indicates that the first exon of *TERT* and the associated bidirectional transcript overlap (Figure 6.2A). RT-PCR verified that at least the first 161 nt of *TERT* exon 1 are shared with TAPAS RNA. It is possible that more of exon 1 is shared between the two genes; however, this could not be confirmed by RT-PCR, likely due to the high GC content of *TERT* exon 1. Additionally, a number of alternatively spliced transcripts were detected, including some skipping exon 2 and others skipping both exons 2 and 3 (Figure 6.2B).

There is a small open reading frame (ORF) (258 nt) that spans exons 4 and 5 and two longer ORFs, of 408 and 375 nts, present in exon 7. However, one of these ORFs is located within the unverified region at the 3' end of the transcript, beyond the main transcription termination sites discussed above. Further, we observed that exon 7 is

poorly conserved between most avian species (*data not shown*). Moreover, no protein domain homology was observed in any region of the transcript (Marchler-Bauer et al., 2014), implicating this transcript as a putative long non-coding RNA (lncRNA).

The recent release of the *Gallus gallus* 5 whole-genome assembly predicts an antisense transcript upstream of *TERT* (LOC107052651). The predicted transcript variant (XR_001465267.1) corresponds to exons 2 through 7 of TAPAS RNA (Figure 6.2A). This variant contains 643 nt more of exon 7 and does not share any sequence with *TERT* exon 1, unlike the transcript reported here. Another transcript variant with retention of an intron between our exons 2 and 3 is also predicted (XR_001465266.1). Further, the NCBI eukaryotic gene prediction tool, Gnomon**,** annotates the predicted transcript as a lncRNA (Accession No. NC_006089).




**<u>Figure 6.2</u>: Schematic of TAPAS gene** (A) Representative Bedgraph from normal bursa tissue shows normalized transcription coverage. Coverage on the plus and minus strands is colored green and blue, respectively. The primary transcript observed by RNA-seq in normal chicken bursa is approximately 9 kb. The principal form of the spliced transcript is 3.6 kb and contains 7 exons. Confirmed region of shared exon 1 is depicted in green-blue stripes. A portion of exon 7 that could not be verified by RT-PCR is indicated in light blue. Transcripts confirmed by RT-PCR are 2-3 kb. The two main 3' ends identified are indicated by vertical orange lines and are located at nucleotide 1051 and 1114 of exon 7. (B) Multiple alternatively spliced variants of TAPAS RNA were also identified in normal bursa by RT-PCR.

*TAPAS RNA expression is elevated in tumors with integrations in the TERT promoter*

The predominance of proviral integrations in the opposite orientation of *TERT,* as well as the identification of a bidirectional transcript, suggested that the integrations might also be driving increased expression of TAPAS RNA. To test this hypothesis, we performed quantitative reverse-transcription PCR (qRT-PCR) to determine TAPAS RNA expression levels in tumors containing integrations in the *TERT* promoter region (Figure 6.3A). Normal liver has 148- and 5-fold less expression than normal bursa for TAPAS and *TERT* RNA, respectively. Compared to normal liver, tumors with integrations in the *TERT* promoter had significantly increased expression of the TAPAS RNA. Expression of the bidirectional TAPAS RNA was up-regulated approximately 250- to 3858-fold relative to normal liver tissue. In contrast, TERT was up-regulated 4- to 42-fold relative to normal liver tissue (Figure 6.3B). This suggests that the observed integrations in the *TERT* promoter are driving expression of a bidirectional lncRNA. These findings were also confirmed by RNA-seq analysis of liver tumors C6, C7 and D2 (*data not shown*).

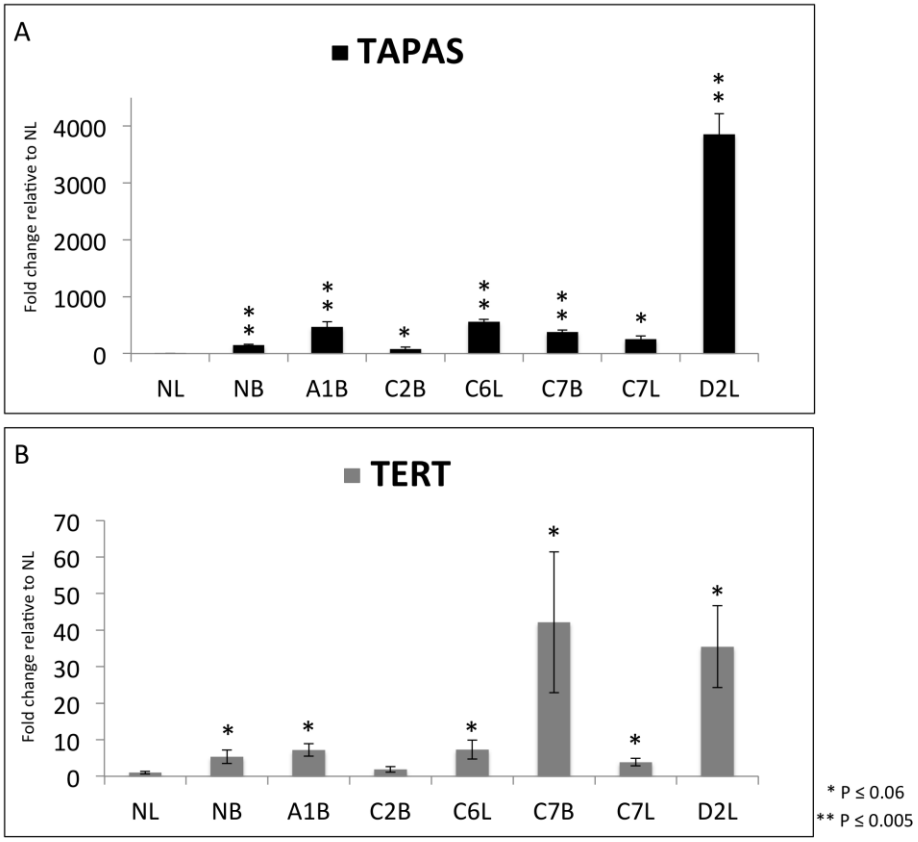



**Figure 6.3: Expression of TAPAS RNA and *TERT* in ALV-induced B-cell lymphomas.** qRT-PCR was performed to determine (A) TAPAS RNA and (B) TERT expression in seven ALV induced tumors as well as normal bursa (NB) and normal liver (NL) controls. Expression of both transcripts is significantly higher in 5 of the 6 tumors as compared to NL. *p-values* are representative of Bonferroni correction for multiple comparisons.

*Viral transcripts splice into exon 4 of TAPAS RNA*

Retroviruses can induce overexpression of host genes by multiple mechanisms (1,44). For instance, insertion of viral enhancer elements in the vicinity of host gene promoters can induce overexpression (Beemon and Rosenberg, 2012). Alternatively, the viral promoter can drive expression of the host gene directly, if both are in the same orientation, by readthrough of the viral poly(A) site (Kanter et al., 1988). If the promoter in the viral 5' LTR is driving expression, the viral RNA transcript can splice via the *gag* splice donor into the cellular mRNA (Kanter et al., 1988). Alternatively, if the promoter in the 3' LTR is used, read-through into the adjacent host genomic region occurs (Hayward et al., 1981). To determine the mechanism by which proviral integrations are affecting TAPAS RNA expression, we analyzed metastasized tumors with integrations in the same transcriptional orientation as the TAPAS RNA. We performed RT-PCR using LTR specific primers and primers within the TAPAS RNA exons to obtain and sequence viral TAPAS RNA fused transcripts (Figure 6.4A).

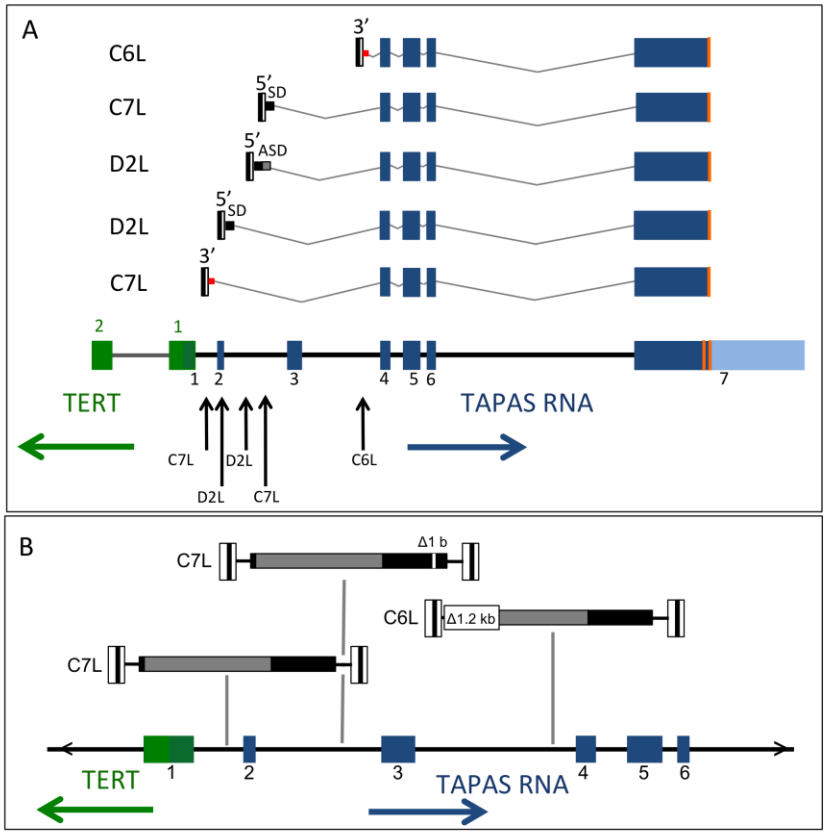
Proviruses in tumor D2L use the 5' LTR to drive expression of TAPAS RNA. One splice variant used the canonical 5' viral splice donor site in *gag* (nucleotide 398). Transcripts were also detected in which an alternative splice donor site in the viral *gag* gene (nucleotide 857) spliced into exon 4 of TAPAS RNA. In tumor C7L, a provirus integrated in intron 2 also spliced into exon 4 of TAPAS RNA from the canonical 5' viral splice donor site.

Alternatively, transcripts in which the viral 3' LTR serves as the promoter were observed in tumor C6L. These transcripts contained 63 nucleotides of host DNA adjacent to the 3' LTR. It appears that a cryptic splice donor site present in this intronic region

may be used to splice into the downstream exon 4. Sequencing of the provirus C6L revealed a large deletion that included the viral splice donor site in *gag* (Figure 6.4B); this would prevent its splicing into the TAPAS RNA if it initiated in the 5' LTR. However, this deletion would probably also prevent transcription initiation at the 5' LTR, as previously observed with ALV integrations in *MYC* (Goodenow and Hayward, 1987). A similar 3' LTR initiated transcript was observed from the C7L proviral integration in intron 1 of TAPAS RNA. This variant spliced from a cryptic splice donor site, 28 nucleotides downstream of the provirus, into exon 4 of the TAPAS RNA.

The majority of proviral integrations are located between exon 1 and exon 4 of the TAPAS gene (Figure 6.4A). However, regardless of the integration site location, all of the viral transcripts analyzed invariantly spliced into exon 4 of the TAPAS RNA. For most of the viral transcripts, this means that nearby exons are skipped and splicing is preferentially occurring to exon 4. While the 5' spliced viral leader sequence contains a bit of the ALV *gag* gene, the analyzed chimeric transcripts have a termination codon before the AUG of the ORF in exon 4 of the TAPAS RNA. Thus, no hybrid protein is predicted. Consistent splicing into exon 4 suggests the possible functional importance of this region of the TAPAS RNA.

**Figure 6.4: Viral RNAs splice into exon 4 of TAPAS RNA.** (A) Splicing of viral transcripts was determined by RT-PCR of tumor RNA. All proviruses indicated in this figure are in the opposite transcriptional orientation of *TERT*. Arrows indicate genomic location of proviral integration. Despite the presence of upstream exons, all viral transcripts analyzed splice into exon 4 of the TAPAS RNA from either canonical gag splice donor (SD) or via alternate splice donor (ASD) sites. Read through transcription from 3'LTR is depicted by a red square. (B) Sequence analysis shows mutations in proviruses C6L and one of the C7L integrations. Three integrated proviruses, two in tumor C7L and one in tumor C6L were sequenced. The viral LTRs are depicted with white boxes at termini of viral genome with the U3, R and U5 direct repeats respectively. *gag-pol* are depicted in grey and *env* in black. The 1.2 kb deletion in C6L removes the canonical viral splice donor and induces transcription from the 3'LTR. Exon positions of *TERT* and TAPAS RNA are depicted for reference.


A
C6L
C7L
D2L
D2L
C7L
3′
5′ SD
5′ ASD
5′ SD
3′
2
1
1 2 3 4 5 6 7
TERT
TAPAS RNA
C7L
D2L
D2L
C7L
C6L
B
C7L
Δ1 b
C6L
Δ1.2 kb
C7L
TERT
1 2 3 4 5 6
TAPAS RNA

*TAPAS RNA is expressed in normal chicken tissues and during development*

To further characterize expression of the TAPAS RNA, we analyzed publicly available RNA-seq data sets from the SRA database (Leinonen et al., 2011). TAPAS RNA and *TERT* expression was measured in various chicken tissues of an 18 day embryo. In addition, transcriptome of total embryos were analyzed at different time points up to 12 days of development (SRA Accession no. ERX697750 and DRX001564 respectively) (Leinonen et al., 2011). Quantification of transcript expression was determined by calculating fragment count normalized to transcript length and total number of reads (FPKM). This analysis showed that the TAPAS RNA is expressed in some normal chick embryo tissues. TAPAS RNA is expressed at particularly high levels in bursa, testes and kidney and is undetectable in muscle and heart tissue (Figure 6.5A). TAPAS RNA expression was higher than *TERT* in bursa and testes but more comparable in other tissues and in total early embryos.

Furthermore, both TAPAS RNA and *TERT* expression is elevated early in chick development and progressively decreases with time (Figure 6.5B). Thus, the tissue specific and developmental expression of TAPAS RNA correlates with *TERT* expression. This suggests that the TAPAS RNA may have a role in regulating *TERT* expression or that the expression of the two transcripts is co-regulated.

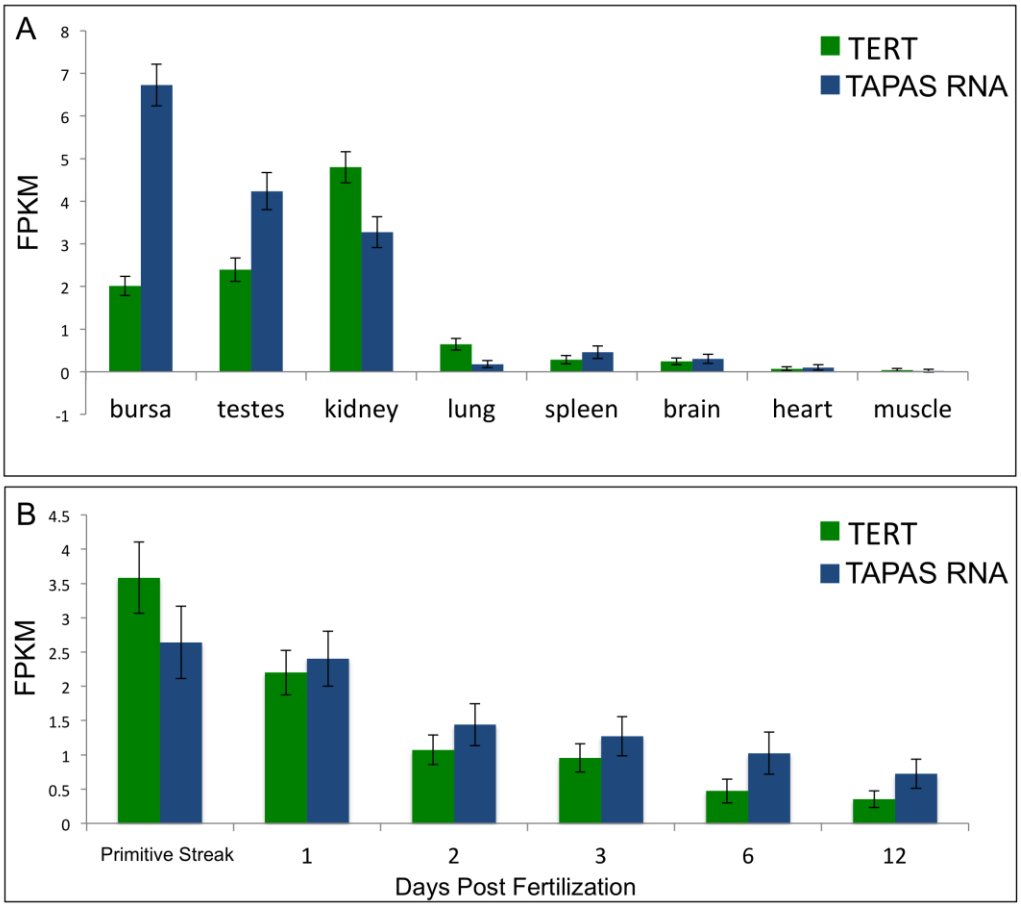



**Figure 6.5: *TERT* and TAPAS RNA are expressed at comparable levels in adult tissues and during chick development.** (A) RNA-seq data of expression of TAPAS RNA in normal 18 day chicken embryo tissues. TAPAS RNA is expressed in various tissues at levels similar to *TERT*. In bursa, TAPAS RNA expression is approximately 3-fold higher than *TERT*. (Pearson correlation coefficient = 0.66, $p$ = 0.07). (B) The expression of *TERT* and TAPAS RNA also seem to be correlated throughout development. (Pearson correlation coefficient = 0.92, $p$ = 0.001)

*TAPAS RNA is conserved*

To determine if the novel TAPAS RNA is conserved, we performed phylogenetic analysis in multiple avian species. Exons 4, 5 and 6 were used for the analysis because this region was found in all alternatively spliced transcripts, as well as in all viral TAPAS RNA fused transcripts, suggesting it may have functional importance. Regions homologous to the TAPAS gene exons 4-6 and intervening introns, were identified in various avian genomes by BlastN, and the sequences were aligned by ClustalX. It was observed that this region is highly conserved at the sequence level in many, but not all, avian lineages (Figure 6.6). Additionally, based on transcriptome analysis, there exists a predicted lncRNA in the TERT promoter region in the most recent genome assemblies for chicken (LOC107052651), turkey (LOC104910189) and Japanese quail (LOC107309454) (*data not shown*). The chicken sequence shared 95% identity with the corresponding genomic region of the closely related black grouse (*Lyrurus tetrix*), and 72-76% identity with the genomes of various neoavian lineages. The turkey and chicken sequences share limited similarities in exons 4-6, but have significant sequence homology in exon 7. The only perching bird species in which this region was conserved was the most basal species of New Zealand wren (*Acanthisitta chloris*) suggesting the possibility that these sequences underwent rapid evolutionary changes in Passseriformes. The ORF in exons 4 and 5 was not conserved in most birds (*data not shown*). Further, we did not find any regions homologous to the chicken TAPAS RNA in mammalian genomes at the sequence level. Exon 6 was the most conserved in avian species with 72-96% identity. The splice donor sites of exon 5 as well as the donor and acceptor sites of exon 6 are perfectly conserved in all species analyzed (*data not shown*).

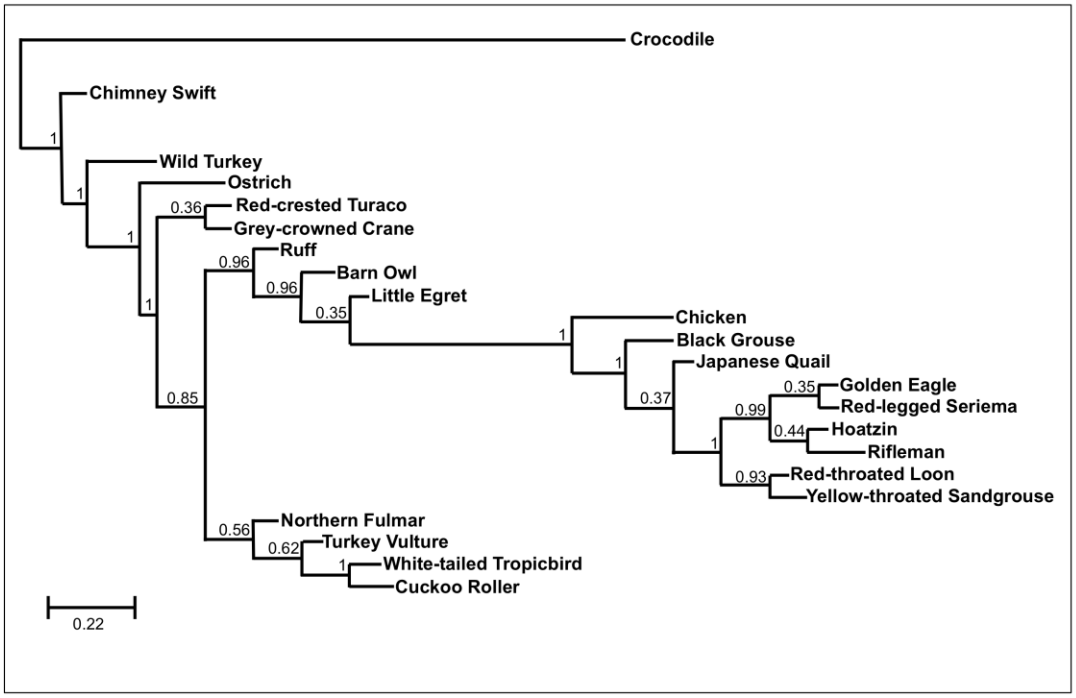



**Figure 6.6: TAPAS gene is conserved in avian species.** Phylogenetic analysis of exons 4-6 and the intervening introns of the TAPAS gene, in several different avian species. Crocodile is depicted as an out-group. Turkey was more conserved in exon 7 than in exons 4-6.

*TAPAS RNA regulates TERT expression*

We reasoned based on the correlation of *TERT* and *TAPAS* expression in tissue and throughout development, that *TAPAS* might be regulating *TERT*. Recruiting chromatin remodelers is a commonly reported function of lncRNAs and would provide a mechanism by which *TAPAS* could be regulating *TERT*. To answer this, we cloned shRNA constructs targeting *TAPAS* into a viral vector. Transfection of chick embryonic fibroblasts (CEF) with the shRNA construct against *TAPAS* generated a robust knockout of approximately 5-fold. In these cells, TERT expression also decreased significantly by 10-fold (Figure 6.7). This provides evidence that *TAPAS* could potentially play a role in regulating *TERT* expression.

We believe this regulation occurs in *cis*- because when the viral fusion truncated TAPAS transcript is overexpressed in *trans*-, there is no significant effect on *TERT* expression (Figure 6.8A). Neither the overexpression nor knockout of *TAPAS* expression caused any significant changes in cell proliferation, resistance to apoptosis, or migration. Interestingly, when *TAPAS* exons 4-7 are overexpressed in *trans* in CEF cells, there appears to be a negative effect on immortalization suggesting that *TAPAS* may promote senescence (Figure 6.8B). However, because *TERT* expression is not affected, the mechanism behind this observed senescence is unclear. Further, this promotion of senescence was not observed when *TAPAS* was overexpressed in DT40 cells.

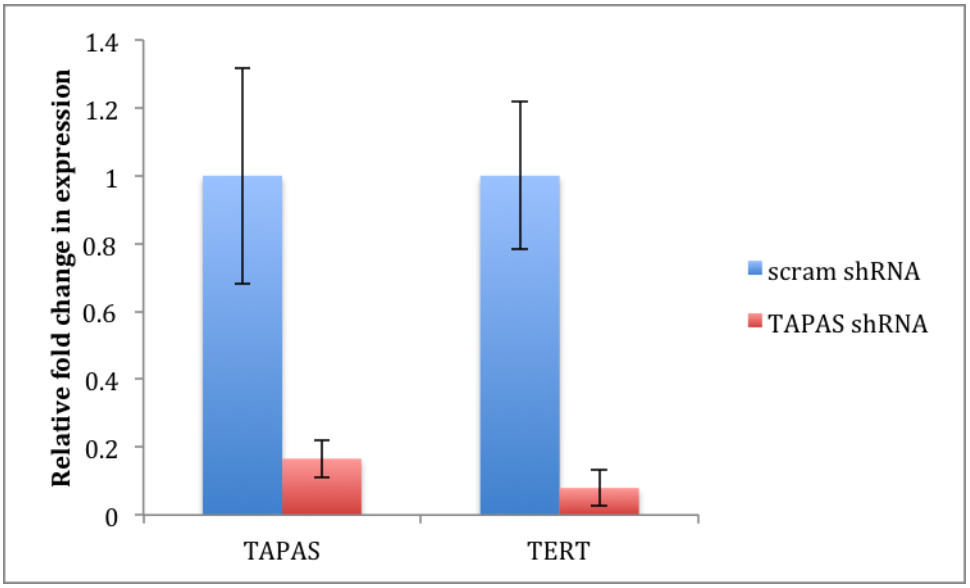
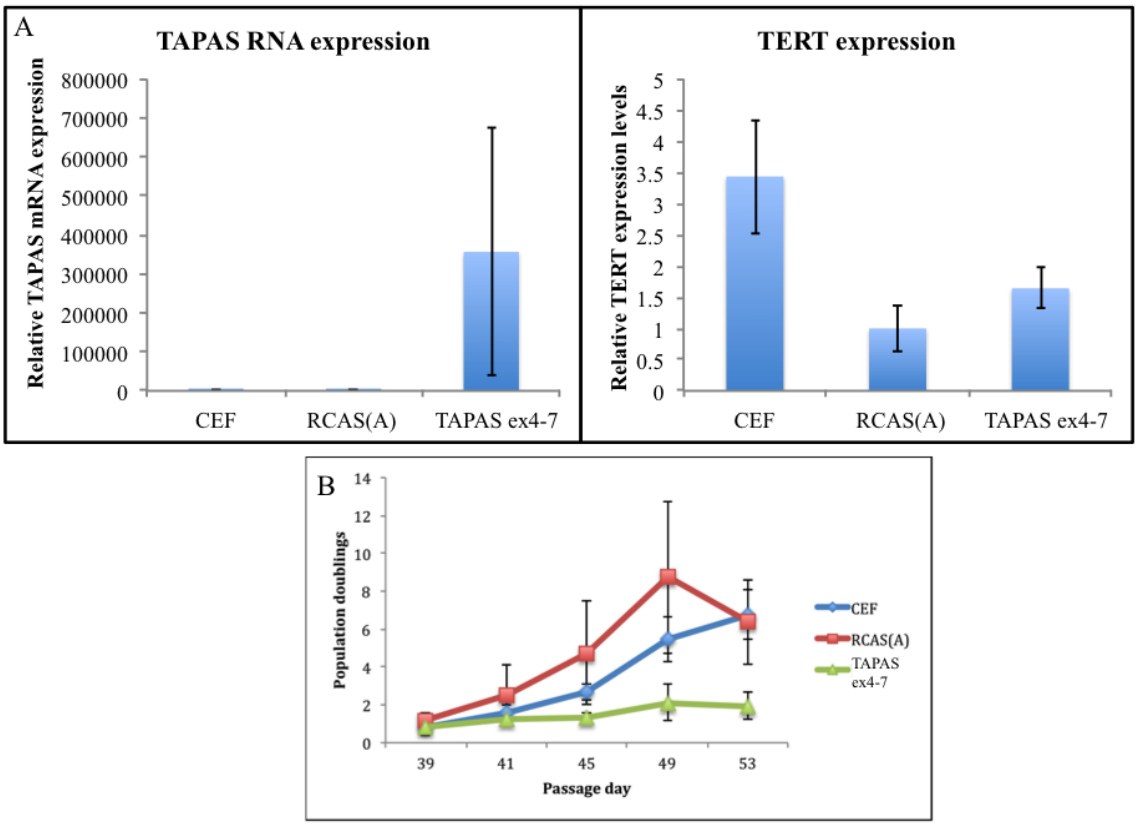



**Figure 6.7: TAPAS RNA knockdown affects *TERT* expression**. (A) shRNA transfection in chick embryo fibroblasts (CEF) decreased *TAPAS* expression by approximately 5-fold relative to a scrambled control. *TERT* mRNA expression decreased 10-fold (n=2).




**<u>Figure 6.8</u>: Overexpression of TAPAS exons 4-7 does not affect TERT expression but does promote senescence in chick embryo fibroblasts.** *TAPAS* exons 4-7 cDNA was cloned into RCAS(A) viral vector to mimic the truncated viral fusion transcript being overexpressed in tumors. Cells were infected with either the RCAS(A) viral vector carrying *TAPAS* or an empty RCAS(A) vector as a control. TAPAS was successfully overexpressed approximately 350,000-fold relative to uninfected CEF cells. (A) *TERT* mRNA expression in cells overexpressing *TAPAS* RNA was not elevated above uninfected CEF cells. (B) Immortalization assay. CEF cells typically senesce after 30 days. Shown is a growth curve measured after this time point, from days 39-53. The average of two duplicate experiments is shown.

*Humans have a similar TAPAS RNA transcript that regulates TERT expression*

To determine if humans have a similar *TAPAS* RNA transcript, we analyzed publicly available RNA-seq data from ENCODE. We found a region upstream of the TERT promoter that is transcribed in a number of cell lines including GM12878, a human B-cell line, and HepG2, a hepatocellular carcinoma line. Similar to chickens, this transcript is antisense. We further analyzed CAGE (Cap Analysis of Gene Expression) data to determine the 5' end of the transcript. CAGE involves pulling down expressed mRNAs by the 5' cap allowing for sequencing and identification of the transcription start site of individual mRNAs. From these datasets, we detected a distinct peak for this transcript that begins about 170 nt upstream of the *TERT* TSS (Figure 6.9). We verified this transcript in various cell lines, such as HEK293T and HeLa, using RT-PCR. We were able to verify a 1.6 kb transcript that is unspliced. The human transcript shares no sequence homology to the chicken transcript. It also has no significant ORFs indicating that it is also a lncRNA.

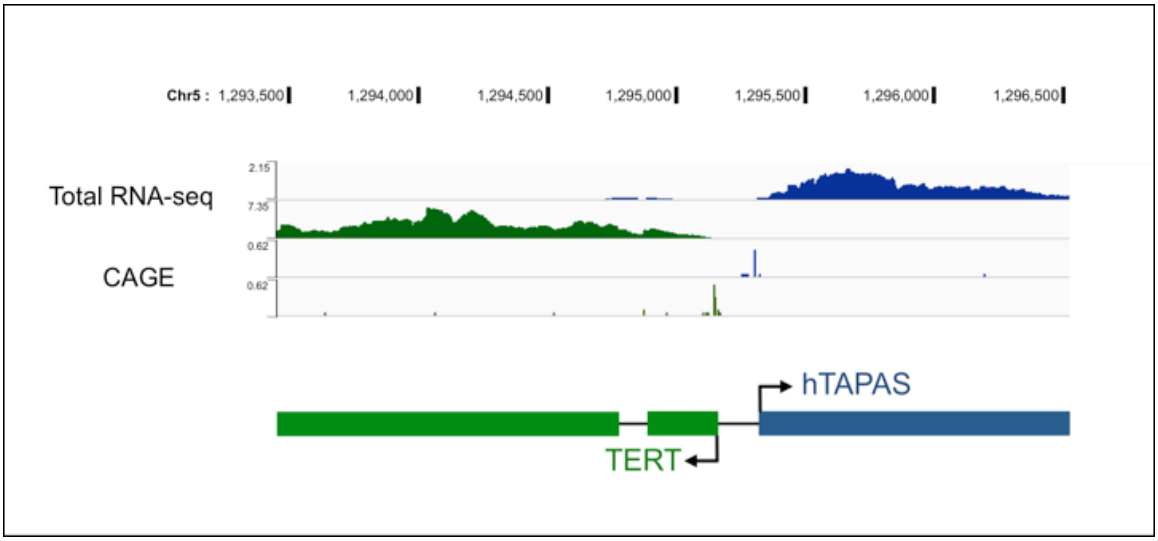
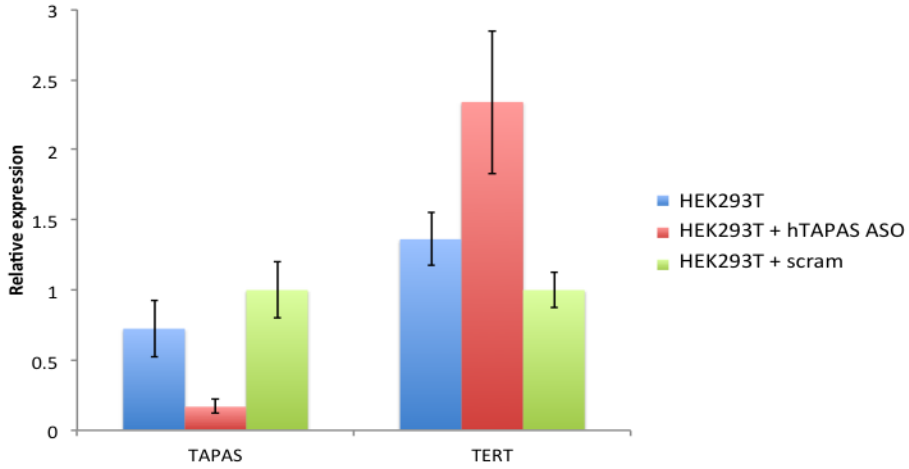



**Figure 6.9: A similar antisense lncRNA transcript can be detected in human cells.** Total RNA-seq data from GM12878 cell lines was used to determine transcription in the region. Transcription is shown as a BedGraph in the top two tracks. Reads shown in green map to the *TERT* transcript. Reads shown in blue map in the opposite transcriptional orientation of *TERT*. The bottom two tracks show CAGE data from the same GM12878 cell line. Distinct peaks can be seen at the 5' end of *TERT* as well as the putative *hTAPAS* transcript. A schematic of the *hTAPAS* gene relative to TERT is shown.

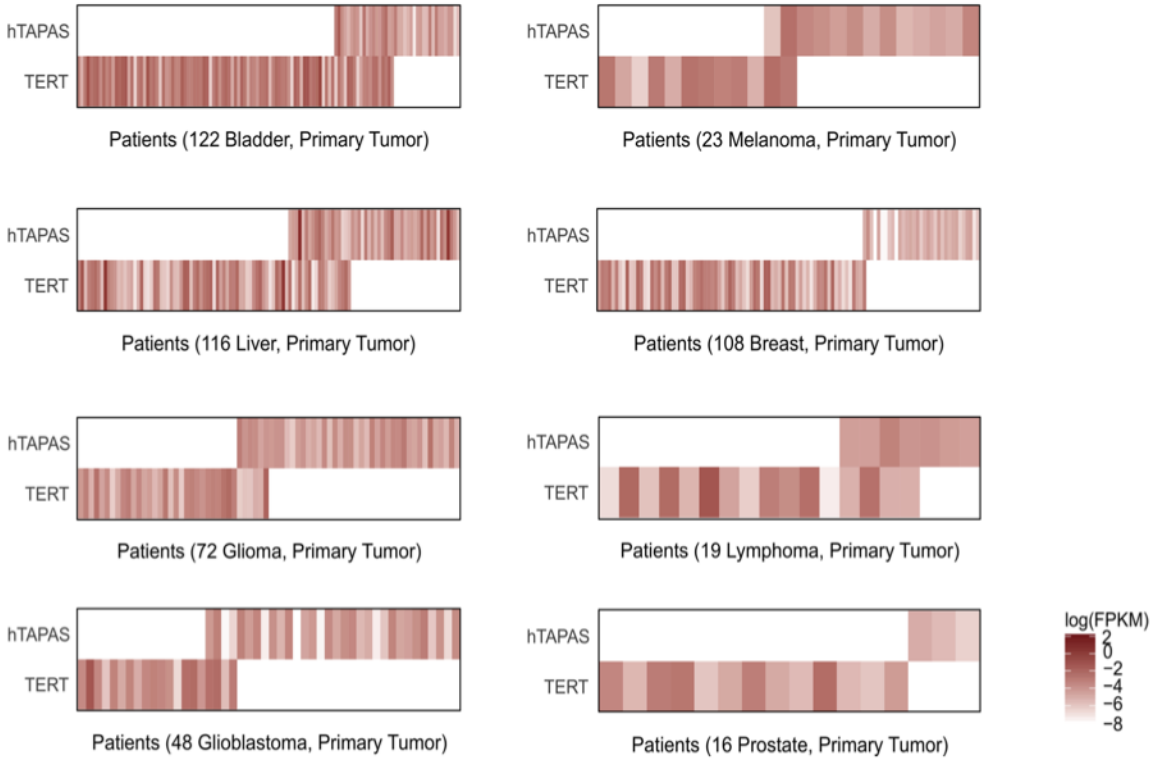
To determine if the putative *hTAPAS* transcript has a similar role in regulating TERT expression, we made use of antisense oligonucleotides targeted against human *TAPAS*. *TAPAS* mRNA levels decline approximately 10-fold relative to a scrambled control when treated with the targeted antisense oligonucleotide. *TERT* expression consequently increased 2.2-fold relative to scrambled indicating that *TAPAS* may similarly be playing a role in regulating *TERT* mRNA expression (Figure 6.10A). To determine if *hTAPAS* is involved in cancer, we analyzed patient transcriptome data from various cancers from The Cancer Genome Atlas (TCGA). Interestingly, we found that *hTERT* and *hTAPAS* are typically not co-expressed in agreement with *in vitro* knockdown data that *hTAPAS* may be negatively regulating *hTERT* expression (Figure 6.10B).

**Figure 6.10: *hTAPAS* may regulate *hTERT* expression**. (A) An antisense oligonucleotide targeting *hTAPAS* was used to knockdown *hTAPAS* levels approximately 10-fold relative to a scrambled control. *TERT* expression subsequently increased approximately 2-fold ($n=3$, $p<0.05$). (B) *hTERT* and *hTAPAS* expression levels in patients with various cancers were determined from RNA-seq data from TCGA. Expression of *hTERT* and *hTAPAS* in patients expressing either of these genes is shown as a heat map with higher expression (FPKM) corresponding to darker red color and lower expression shown as lighter color. *hTAPAS* expression is shown on the top track while TERT expression is shown on the bottom track. Data is separated by cancer type. (M. Freeberg and S. Malhotra)

A


Relative expression
3
2.5
2
1.5
1
0.5
0
TAPAS
TERT
HEK293T
HEK293T + hTAPAS ASO
HEK293T + scram

B


hTAPAS
TERT
Patients (122 Bladder, Primary Tumor)
Patients (23 Melanoma, Primary Tumor)
Patients (116 Liver, Primary Tumor)
Patients (108 Breast, Primary Tumor)
Patients (72 Glioma, Primary Tumor)
Patients (19 Lymphoma, Primary Tumor)
Patients (48 Glioblastoma, Primary Tumor)
Patients (16 Prostate, Primary Tumor)
log(FPKM)
2
0
−2
−4
−6
−8

## 6.3 Discussion

In order to better understand tumor pathogenesis, we analyzed the distribution of ALV integration in chicken B cell lymphomas using a high throughput sequencing strategy. We identified numerous clonally expanded proviral integrations in the *TERT* promoter region suggesting these integrations may promote tumorigenesis. Previous studies have also reported proviral integrations of ALV, hepatitis B and papilloma virus in the *TERT* promoter region that induced elevated *TERT* expression, with similar implications in tumorigenesis (Ferber et al., 2003; Li et al., 2014; Yang et al., 2007b). However, in our work we show that the proviral integrations in our tumors are antisense to *TERT* and are driving the over-expression of a novel lncRNA, which we've named *TAPAS* (*TERT* antisense promoter-associated) RNA. Retroviral activation of a birectional promoter-associated lncRNA has not been previously reported. The prevalence of integrations in the same orientation to the *TAPAS* RNA in multiple tumors suggests that over-expression of this transcript may have made cells predisposed to oncogenic transformation and proliferation. We also find evidence for a similar upstream antisense transcript in human cell lines.

The detected *TERT* promoter associated transcripts in chickens and humans vary significantly. No conservation was detected at the sequence level. The chicken transcript contains 7 exons and is alternatively spliced whereas only one, unspliced exon has been detected in the human transcript. Similar to the chicken transcript however, the human transcript has no significant ORFs and no protein domain homology indicating that it is also a long noncoding RNA.

Many lncRNAs are known to play a role in transcriptional regulation. For example, *XIST* acts in cis to recruit the polycomb repressive complex 2 (PRC2) to chromosome X causing gene silencing (Zhao et al., 2008). *HOTAIR* on the other hand acts in trans to repress the expression of genes in the HoxD gene cluster (Gupta et al., 2010). It has been proposed that antisense lncRNAs transcribed from bidirectional promoters may be involved in regulation of the associated sense transcripts (Wakano et al., 2012; Wei et al., 2011). Such an arrangement may allow for tighter transcriptional regulation. Based on the correlation between *TERT* and *TAPAS* RNA expression in adult tissues, we hypothesized that TAPAS may be regulating *TERT* expression.

Consistent with this idea, knockdown of *TAPAS* lncRNA expression in *cis*-affected *TERT* mRNA expression. In CEF cells, knockdown of *TAPAS* caused a decrease in *TERT* expression, while in humans, *hTAPAS* knockdown caused a subsequent increase in *TERT* expression. Thus, while it seems that *TAPAS* is affecting *TERT* expression, it is unclear how it is doing so. However, we do believe this regulation is likely occurring in *cis*- as episomal over-expression of the truncated viral fusion *TAPAS* transcript had no impact on *TERT* expression. It is also possible that *TERT* and *TAPAS* expression are only correlated due to mutual regulation by the bidirectional promoter. This opens the possibility that *TAPAS* RNA may have an independent function apart from that of *TERT* regulation.

Given the important role of lncRNAs in transcriptional regulation, it comes as no surprise that many lncRNAs have been implicated in cancer and disease. With the advent of deep sequencing, an increasing number of lncRNAs have been identified that have differential expression in cancer tissues. Many such novel lncRNAs are associated with

tumorigenesis and cancer pathogenesis. For example, *BIC* was first identified as a non-coding RNA up regulated in ALV induced lymphomas and was later found to be the precursor of the oncogenic microRNA *mir-155* (Clurman and Hayward, 1989). Other well-studied lncRNAs, such as HOTAIR, MALAT1, PCAT-1, PCGEM1, TUC338 were reported as oncogenes, while GAS5, MEG3 and PTENP1 were reported as tumor suppressors (Braconi et al., 2010; Gibb et al., 2011; Gupta et al., 2010; Gutschner et al., 2013; Huarte et al., 2010; Mourtada-Maarabouni et al., 2009; Poliseno et al., 2010; Prensner et al., 2011; Zhang et al., 2010). Since lncRNA functions range from cell growth to cancer development, they represent important biological players that merit further research.

We have identified novel antisense lncRNAs transcribed from the *TERT* promoter in chickens and humans. We show that these lncRNAs are up regulated in cancer and implicated in tumorigenesis. We also show that they may play a role in regulating *TERT* mRNA expression. Further characterization of the structural and functional motifs of these lncRNAs is required to better understand its mechanistic and functional role in cancer signaling. This will help elucidate possible gene regulatory mechanisms of lncRNAs, and will provide insight into the role lncRNAs play in cancer pathogenesis.

# Chapter 7 – Future directions

## 7.1 Regulation of ALV integration

We provide here strong evidence that the FACT complex directly binds ALV integrase and regulates ALV integration efficiency *in vitro* and *in vivo* (Chapter 3). However, the evidence that the FACT complex is also playing a role in targeting of ALV integration is less robust. We do observe that knockout of the FACT complex causes a significantly decreased frequency of integration in the proximity of the transcription start site (TSS). Data from FACT complex chromatin binding in human cells indicates that the FACT complex is enriched at the TSS, thus this may indicate that the FACT complex is recruiting the ALV integration complex to its chromatin binding sites. Performing integration mapping in wild type human cells as well as in cells with functional FACT complex and cells in which the FACT complex has been depleted would allow correlation of integration sites with FACT complex ChIP-seq data to determine if FACT binding sites overlap with integration sites. This would allow the question of the mechanism of FACT targeting to be more thoroughly addressed.

Further, integration of ALV into satellite sequences also correlates with FACT complex levels. In the absence of FACT complex, there appears to be significantly more integration into satellite sequences. With the limited amount of information available on the chicken genome it was not possible to determine what these satellite sequences were. We hypothesize that these satellites might be a general marker of heterochromatin and that the absence of the FACT complex results in more integration into heterochromatic regions. Again, experiments in human cells in which information on heterochromatin and euchromatin are available would be enlightening. Furthermore, correlating chromatin

modifications with integration site location may reveal more differences in integration targeting that we are unable to detect in chicken cells.

Our data seems to suggest that the BET proteins may also play a secondary role in regulating and/or targeting ALV integration. However, this regulation seems to be opposite in nature to that observed for the FACT complex. While the FACT complex promotes ALV integration efficiency, BET proteins inhibit integration. We believe that the role of BET proteins is secondary due to the primary effect of the FACT knockout on ALV integration efficiency, in addition to the observation that BET protein inhibition affects ALV integration pattern more appreciably in the absence of functional FACT complex.

Binding of BET proteins to the ALV integrase protein needs to be evaluated to determine if the binding is direct. Once this is determined, competition between FACT complex and BET protein binding can be assessed. *In vitro* integration assays of BET proteins alone as well as in conjunction with the FACT complex would also be useful in developing a model for ALV integration regulation. Evaluating the integration pattern of ALV in human cells would also be helpful. The chicken genome is poorly annotated and there is little additional data on histone modifications or transcription factor binding sites. Having integration data in human cells in which FACT and/or BET proteins have been depleted or inhibited would allow correlation of integration pattern with many additional genomic features.

The fact that ALV integration is relatively random may also make it difficult to see differences in integration pattern when host cell factors are depleted. To show that

certain host cell factors do indeed play a role in targeting via a bimodal tether model, chimeric factors could be constructed to achieve retargeting.

Lastly, our screen to identify integrase binding partners is limited. We make the assumption that the host cell factor responsible for regulation or targeting of ALV binds the integrase protein alone. *In vivo*, the integrase protein is not present in isolation. Instead it is complexed with viral DNA, as well as other viral proteins such as nucleocapsid and capsid, within the pre-integration complex (PIC). Thus, it is possible that our screen is not detecting all cooperating host cell factors. To further complicate the matter, the integrase protein is also extensively modified post-translationally which could alter the repertoire of host cell proteins that it is capable of binding (Cereseto et al., 2005; Zheng and Yao, 2013). Thus, modifying the screen to be more similar to the in vivo context of the integration complex may be able to reveal more regulatory host cell factors.

## 7.2 Function of CTDSPL and CTDSPL2 in oncogenesis

While previous studies have characterized CTDSPL as a tumor suppressor gene, our studies indicate to the contrary that both CTDSPL and CTDSPL2 have oncogenic properties. In the tumors, ALV integrations into these genes generate truncated viral fusion transcripts. It is unclear exactly what the truncation is doing to alter the function of the proteins and why these truncated products have been selected for in tumors. Expression of truncated constructs uniquely promotes immortalization, a function not observed when the full-length proteins are expressed alone. The portions of the proteins that are truncated contain an intrinsically disordered region that could potentially play a role in mediating substrate interaction.

CTDSPL has been reported to act as a phosphatase on pRb (retinoblastoma protein), a well-studied tumor suppressor gene. By dephosphorylating pRb, CTDSPL keeps it in an active form. We hypothesize that the truncated proteins are no longer able to interact with pRb and thus pRb becomes inactivated allowing for avoidance of senescence. To test this hypothesis, we could analyze phosphorylation status of pRb *in vivo* in the presence of the truncated proteins as opposed to the full-length proteins. Immunoprecipitation of full length and truncated proteins from cell extracts could also be used to determine differential binding partners which could help elucidate the function of the truncated protein.

## 7.3 Function of chicken and human TAPAS RNA

We show here preliminary data on the function of TAPAS RNA in regulating TERT expression. We find that chicken and human TAPAS seem to regulate TERT mRNA expression in opposite directions. It is unclear whether this is a species specific or cell-type specific effect. Of note, the chicken knockdown experiments were performed in a primary cell line whereas the human knockdown experiments were done in an immortalized cell line (HEK293T). Preliminary data not shown here indicates that in a B-cell lymphoma derived chicken cell line (DT40s), knockdown of TAPAS may cause an increase in TERT expression similar to that observed in the HEK293T cell line. Thus, it is possible that TAPAS may function differently in primary vs. immortalized or cancer-derived cells. It would be useful to perform TAPAS RNA knockdown experiments in more cell types of each species to determine whether the effect is cell-type specific. The mechanism by which TAPAS RNA is regulating TERT expression is not clear. We hypothesize that as a lncRNA, TAPAS might be regulating TERT by regulating

chromatin remodeling at the adjacent TERT promoter. Moreover, to narrow down the function of TAPAS RNA, we could perform RNA immunoprecipitation from cell extracts to determine what TAPAS is binding. This could determine what, if any, chromatin remodelers TAPAS is binding and potentially recruiting. We could then subsequently look at histone modifications and/or DNA methylation at the TERT promoter in the presence and absence of TAPAS RNA expression to determine if TAPAS does indeed modify the chromatin status.

Further, we have only explored the hypothesis that TAPAS regulates TERT mRNA expression. It is possible that TAPAS RNA could be regulating other cellular genes or processes. To determine if this is the case, we could overexpress or knockdown TAPAS RNA and perform transcriptome profiling to look for any differential expression.

# Chapter 8 – Materials and methods

*Cell culture and viruses*

DT40 cells were cultured in Dulbecco's modified eagle medium (DMEM; Thermo Fisher Scientific), 10% fetal calf serum, 5% chicken serum, 5% tryptose phosphate and 1% antibiotic at 37°C and 5% $CO_2$ (Winding and Berchtold, 2001). CEF cells were cultured in media 199 (Thermo Fisher Scientific) supplemented with 2% tryptose phosphate, 1% fetal calf serum, 1% chicken serum and 1% antibiotic/antimycotic at 39°C and 5% $CO_2$. HEK293T cells were cultured in Dulbecco's modified eagle medium, 10% FBS and 1% antibiotic at 37°C and 5% $CO_2$. HEK293T cells were cultured in Dulbecco's modified eagle medium (Thermo Fisher Scientific) supplemented with 10% fetal bovine serum at 37°C and 5% $CO_2$.

SSRP1 conditional knockout cells ($SSRP1^{-/-}$ + SSRP1) were obtained as a gift from Dr.Takemi Enomoto, Tohoku University (Abe et al., 2011). This cell line was generated by knocking out both endogenous *SSRP1* loci in an otherwise wild type DT40 background. A Flag-tagged wild type copy of the chicken *SSRP1* gene was introduced under the control of a tet-repressible promoter. To induce knockout, cells were pre-treated with 1 μg/mL doxycycline for 24 hours before infection. Knockout was verified by qRT-PCR or Western blot prior to infection using a Flag antibody (8146, Cell Signaling Technology).

ALV was generated by transfecting CEF cells with RCASBP(B), RCASBP(C) or RCASBP(C)-eGFP plasmid (Hughes, 2004) using electroporation. Viral supernatant was collected and filtered through a 0.22 micron filter. To generate MLV and HIV-1 pseudotyped with vesicular stomatitis virus G glycoprotein (VSV-G), NIH-3T3 or HEK293T cells were co-transfected using lipofectamine (Thermo Fisher) with pMD.G

(VSV-G envelope plasmid) (Burns et al., 1993) and either MLV (pNCS) (Gao and Goff, 1998) or HIV plasmid (pNL4-3ΔE-GFP) (Zhang et al., 2004) respectively. Viral supernatant was collected after 48 hours, filtered through a 0.22 micron filter and concentrated by PEG precipitation (10% PEG 8000) (Cepko et al., 2001).

*Reverse transcriptase assay for viral detection*

Viral supernatant was collected from infected cells by removing cellular media and centrifuging to remove cells. Viral supernatant was then mixed with a cellular RNA template, random hexamer primer, dNTPs, and RT buffer (20 mM $MgCl_2$, 0.2% Triton-X 100, 2 mM EDTA, 100 mM Tris, pH 8.0). The mixture was incubated for 1 hour at 37°C. Samples were then diluted 2-fold and 1 uL of reaction is used in a standard SYBR Green qPCR reaction with primers against a housekeeping gene.

*siRNA, shRNA and antisense oligonucleotide mediated gene knockdown*

siRNA against NCL and UBTF were used to knockdown these genes in HEK293T cells. siRNA sequences can be found in Appendix 1. HEK293T cells were transfected using FuGene 6 (Promega). chTAPAS was knocked down using cloned shRNA, the sequences of which can be found in Appendix 1. shRNA were cloned into RCAS(A) vector. Virus was produced in CEF cells and subsequently used to transmit the shRNA to experimental cells. hTAPAS was knocked down in HEK293T cells using antisense oligonucleotides delivered using FuGene 6 for transfection.

*Nucleic acid extraction*

Genomic DNA was isolated using a DNeasy Blood & Tissue kit (Qiagen). RNA was extracted from tissue homogenates using RNA Bee extraction agent (Tel-Test, Inc, Friendswood, TX).

*Integration site mapping high throughput sequencing library preparation and analysis*

Genomic DNA for ALV integration mapping libraries was collected using standard proteinase K digestion followed by phenol-chloroform extraction (Yang et al., 2007b). Libraries were prepared and analyzed using a custom pipeline as described previously (Justice et al., 2015a). Integrations were attributed to the nearest RefSeq gene. To analyze integration pattern into various genomic annotations as well as around TSS we made use of HOMER bioinformatics tools (Heinz et al., 2010). Integration frequency in proximity to CpG islands was calculated using BedWindow (Quinlan and Hall, 2010).

*Transcriptome profiling and analysis*

RNA-Seq libraries were prepared in duplicate using the TruSeq stranded mRNA library kit according to manufacturers directions and sequenced on the Illumina HiSeq platform. Differential gene expression was determined using Cufflinks (Trapnell et al., 2012). Genes with a 2-fold or greater difference in gene expression were considered for further analysis. Gene ontology (GO) analysis was performed using g:Profiler and DAVID (Huang et al., 2009; Reimand et al., 2007). GO terms with a p-value of less than 0.05, after Bonferroni correction for multiple testing, were considered significantly enriched above background.

Additional RNA-Seq data for analysis of tissue distribution and embryonic expression of TAPAS RNA and *TERT* in chickens were downloaded from the public sequence read archive (SRA) database (SRA Accession no. ERX697750 and DRX001564) (Leinonen et al., 2011). Abundance of transcripts (FPKM) was estimated and compared using Cufflinks (Trapnell et al., 2012).

*Recombinant proteins, affinity pull-down and MS-based proteomics to identify integrase binding factors.*

6xHis-tagged HIV-1 IN, GST-tagged HIV-1 IN and 6xHis-tagged MLV IN were purified as previously described (Larue et al., 2014; McKee et al., 2008). Full length GST-tagged ALV IN was made synthetically in pGEX-6P-1 by GenScript and truncated by site directed mutagenesis to add a stop codon at codon 51 and 208 (generating GST-tagged ALV NTD IN and GST-tagged ALV NTD/CCD IN respectively). ALV IN constructs were purified similarly to HIV-1 IN with either a HisTrap HP column followed by Heparin column or Glutathione Sepharose column (all from GE Healthcare). The FACT proteins were purified as described (Winkler et al., 2011).

To identify cellular proteins selectively interacting with ALV IN, in parallel reactions we used recombinant ALV and HIV-1 INs as baits to capture their binding partners from cellular extracts. Affinity pull-down experiments were performed with GST- and 6xHis-tagged proteins using glutathione sepharose 4B and nickel affinity beads (GE Healthcare), respectively. Buffer conditions were 25 mM Tris (pH 8.0), 200 mM NaCl, 0.1% Nonidet P-40, 2 mM β-mercaptoethanol, and 1×complete protease mixture (Roche) or 50mM TrisHCl pH 7.5, 250mM NaCl, 2 mM β-mercaptoethanol, and 1×complete protease mixture (Roche), respectively for nickel and GST beads. Sup-T1, DT40 or 293T nuclear extracts were prepared using the NE-PER Nuclear and Cytoplasmic Kit (Thermo Scientific) and incubated with the prebound beads and the bound proteins were separated by SDS-PAGE. Samples were either subjected to immunoblotting using SSRP1 (ab137034, Abcam) or Spt16 (sc-28734, Santa Cruz) antibodies or analyzed by MS.

For MS experiments, entire lanes were excised, subjected to in-gel trypsin digestion and the resulting peptides were analyzed with capillary-liquid chromatography – tandem MS/MS using a Thermo Finnigan LTQ Orbitrap mass spectrometer equipped with a microspray source (Michrom Bioresources). We performed two sets of pull-downs from nuclear extracts of DT40 and Sup-T1 cells for the MS experiments. Human Sup-T1 cells were used in addition to chicken DT40 cells because the MASCOT search engine allows for peptide mass fingerprinting using Homo sapiens (human) but not chicken taxonomy. This is because the chicken genome is currently not completely structurally and functionally annotated (based on Gene Ontology) and many of the genes have not yet been assigned standard nomenclature. Therefore, to identify the peptides from DT40 cells we used the higher order "bony vertebrate" classification. For both set of pull-downs; unique proteins (those with a spectral count greater than 5) were identified that bound either HIV-1 or ALV IN, compared between cell types and only those that were reproducible unique hits were selected for further analysis. LEDGF/p75 served as a control as it is known to selectively bind HIV-1 IN.

*Integrase binding experiments*

Recombinant protein pull-downs were performed with purified GST-tagged HIV-1 or ALV INs as well as the ALV IN domains (1 μM) prebound to glutathione sepharose 4B beads (GE Healthcare) in 50mM TrisHCl pH 7.5, 250mM NaCl, 2 mM β-mercaptoethanol, and 1×complete protease mixture (Roche). Purified SSRP1, Spt16 or the FACT complex (0.6 μM) were added to the beads, and the bound proteins were separated by SDS-PAGE and visualized by Coomassie staining.

*Homogenous time-resolved fluorescence – in vitro integration assay*

HIV-1 or ALV IN strand transfer activities were assayed using similar homogenous time-resolved fluorescence (HTRF)-based strand transfer assays developed for HIV-1 and MLV IN (Kessl et al., 2012; Sharma et al., 2013). The assays contained 5′-Cy5–labeled viral donor DNA (200 nM) (ALV Don1: /5Cy5/ACGAGCACAGGAGTATGGATGACGACAACATT, ALV Don2: /5Cy5/AATGTTGTCGTCATCCATACTCCTGTGCTCGT), biotin-labeled target DNA (20 nM) (Ace1: ACAGGCCTAGCACGCGTCG/3'Bio/, Ace2: CGACGCGTGCTAGGCCTGT/3'Bio/), purified recombinant His-tagged ALV IN or HIV-1 IN (400 nM) and purified recombinant SSRP1, Spt16, or FACT complex (1 μM) to the respective reactions containing IN, donor (HIV-1 or ALV specific), and target DNA substrates. The strand-transfer products were detected after addition of europium chelate–streptavidin Lance reagent (2 nM; PerkinElmer). The HTRF signal was recorded using a Perkin-Elmer Multimode Enspire plate reader using 314 nm for excitation wavelength and 668 and 620 nm for the wavelength of the acceptor and donor emission, respectively.

*Quantification of mRNA expression levels*

Reverse transcription was performed with Maxima H Reverse Transcriptase (Thermo Fisher Scientific) using an oligo$(dT)_{18}$ and random hexamer primer. Quantitative PCR reactions were performed with the CFX96 Real Time System (Bio-Rad) and prepared using PowerUp SYBR Green Mastermix (ThermoFisher Scientific) according to the manufacturer's protocol on a BioRad C1000 thermocycler / CFX96 Real-Time System. qPCR was performed in triplicate and analyzed using the comparative $C_t$ method (ΔΔCt).

*Quantification of proviral integration and other viral intermediates.*

To quantify proviral integrations, genomic DNA was purified by gel purification. Total DNA was loaded on a 0.5% low melting point agarose gel and run at 100V for 3 hours. The high molecular weight band was then purified using a QiaEx II gel purification kit (Qiagen). The purified gDNA was then subjected to qPCR analysis with viral specific primers (ACATCCTTCTGACCGACCCA, CAATTCTGTCTCATTTGGGAGCAA) or from unpurified DNA using a CR1-*gag* nested PCR approach. In this approach, a PCR reaction was performed using a forward primer in the CR1 repeat element (N(8)ATTCTRTGATTCTRT) and a reverse primer in the *gag* gene of ALV (TAGGTTTTACACGCGGACGA). The product of this reaction was used as a template for a qPCR reaction with viral specific primers located in the LTR (ACCGTTGATTCCCTGACGAC, TGGCCGACCACTATTCCCTA).

2-LTR circles were detected by qPCR using primers spanning the LTR-LTR junction (GACTACGAGCACCTGCATGA, TCTCCTTGTAAGGCATGTTGCT). Plus strand extension products were quantified with *gag* forward and reverse primers (CTTGGGGAGTCCAACTCCAG, AGCCGGGCAACTTCTCTAAA). MLV and HIV-1 integrations were quantified with viral specific primers from gel purified genomic DNA (MLV *gag*: TCAGGTCGGGCCACAAAAAC, ACTAGCTCTGTATCTGGCGGA; HIV-1 *eGFP*: ATCATGGCCGACAAGCAGAA, TCTCGTTGGGGTCTTTGCTC). Relative quantification was performed using the $2^{-\Delta\Delta CT}$ method.

*Tumor induction*

All of the B-cell lymphomas included in this study were rapid-onset lymphomas induced by either wild type (WT) or variants of LR-9 virus infections, in 10 day old

chicken embryos, as described previously (Polony et al., 2003). LR-9 is an ALV subgroup A recombinant virus consisting of *gag*, *pol*, and *env* genes derived from UR2 associated virus and LTRs derived from ring-necked pheasant virus (Simon et al., 1987). Tumors were collected from primary bursal (B) tumors or metastasized liver (L) tumors. A1B was induced by ΔLR-9, with a deletion in the *gag* gene, causing increased splicing to downstream genes (Smith et al., 1997). Tumors C2B, C2L, C6L, C7B and C7L were induced by infection with LR-9 containing a silent mutation, G919A, which induces a higher incidence of rapid-onset lymphomas (Polony et al., 2003), probably due to increased readthrough and splicing to downstream genes (O'Sullivan et al., 2002). Tumor D2L was induced by WT LR-9 (Simon et al., 1987).

*Detection of viral fusion transcripts*

RNA was extracted using RNA-Bee reagent (Tel-Test, Inc.). cDNA was prepared using Maxima H reverse transcriptase with an oligo$(dT)_{18}$ primer (ThermoFisher Scientific). Fusion transcripts were detected by performing PCR with a forward primer in *gag* immediately before the viral splice donor (TCAAGCATGGAAGCCGTCATAAAG) and a reverse primer within the gene of interest (*CTDSPL*: TGAAAATGCAGTGCCTGTGC; *CTDSPL2*: CAGTAAGGTAGTTCGCGGGG, *TAPAS*: CAAATGGCTTGTCTGCATTTTCTTC,

CCAAAGCCACGGCTTCCATGTTAGTATC,

TAAGGTGGAGAATAAGACATAATAATATGAGATGAG).

*Proliferation and apoptosis assay.*

Cells were seeded at 0.8 x $10^6$ cells in a 10 cm dish at time 0. To induce apoptosis, cells were treated with 50 μM $H_2O_2$ as described (Jin et al., 2011). After 48

hours, cells were collected and counted using a BioRad automated cell counter (BioRad TC20) to determine change in cell survival relative to CEFs infected with empty viral vector. Population doublings were calculated from total live cell count at day 2 relative to day 0. Proliferation was then plotted relative to CEFs infected with empty viral vector as a control condition. Significance was assessed using an unpaired t-test.

*Scratch assay.*

A scratch assay was used to detect differences in cell migration as described previously (Liang et al., 2007). Cells were seeded in 6-well plate at 3x105 cells per well and allowed to grow for approximately 24 hours or until cells were 100% confluent. The plate of cells was then scratched with a P200 tip at time 0. Closure of the scratch was monitored via light microscopy for 8 hours. Migration of cells into scratch was quantified using ImageJ (Abramoff et al.).

*Evolutionary analysis of TAPAS RNA*

The sequence encompassing exons 4-6 and intervening introns of chicken TAPAS gene was analyzed via BlastN against the whole genome shotgun database at the NCBI web site (http://blast.ncbi.nlm.nih.gov/). All high quality matches with an E-value lower than $10^{-40}$ were retrieved and further analyzed. Sequences were aligned and plotted via the maximum-likelihood method, by PhyML, utilizing the GTR substitution and aBayes branch support (Guindon et al., 2010).

## Appendix I: Primers and oligonucleotide sequences

| | | |
|---|---|---|
| **SSRP1 F** | GGCTCACCAAGAACATGTCA | **qPCR primer for chicken SSRP1** |
| **SSRP1 R** | AGTCCTGAGCTGGCCTTGTA | **qPCR primer for chicken SSRP1** |
| **SW20** | N(8)ATTCTRTGATTCTRT | **CR1 nested PCR 1** |
| **SW23** | TAGGTTTTACACGCGGACGA | **CR1 nested PCR 1** |
| **SW29** | ACCGTTGATTCCCTGACGAC | **CR1 nested PCR 2** |
| **SW30** | TGGCCGACCACTATTCCCTA | **CR1 nested PCR 2** |
| **SW39** | TCTCCTTGTAAGGCATGTTGCT | **ALV 2LTR qPCR** |
| **SW40** | AACGCCATTTGACCATTCACC | **ALV 2LTR qPCR** |
| **SW87** | ACCTGGGGATTGGTTTTGGG | **ALV PSE qPCR** |
| **SW88** | TGGTTTCTCGATGCACTCCG | **ALV PSE qPCR** |
| **SW53** | GCCACTGTCGTTAGTGGACA | **MLV quantification** |
| **SW54** | AATCTTTAGCCCAGTGCCCC | **MLV quantification** |
| **SW50** | ATCATGGCCGACAAGCAGAA | **HIV-GFP quantification** |
| **SW51** | TCTCGTTGGGGTCTTTGCTC | **HIV-GFP quantification** |
| **SW199** | CACATGGTCCTGCTGGAGTT | **Sequencing primer in RCAS(C)-GFP** |
| **Hs.Ri.NCL.13.1** | AUAUUGAAUUUAAGACAGAAGCUGA | **Human NCL targeting siRNA 1** |
| **Hs.Ri.NCL.13.2** | CCGUGUUGGUUUUGACUGGAUAUTC | **Human NCL targeting siRNA 2** |
| **Hs.Ri.NCL.13.3** | UCUCUUUGUUGGAAACCUAAACUTT | **Human NCL targeting siRNA 3** |
| **Hs.Ri.UBTF.13.1** | GUCAUGUCACUGACCUAUUAAAUTG | **Human UBTF targeting siRNA 1** |
| **Hs.Ri.UBTF.13.2** | AACCAAGAUUCUGUCCAAGAAAUAC | **Human UBTF targeting siRNA 2** |
| **Hs.Ri.UBTF.13.3** | CGUGCAGCAUAUAAAGAGUACAUCT | **Human UBTF targeting siRNA 3** |
| **SW293** | TTGAGGGCAGAGCAATCAGG | **Human NCL qPCR** |
| **SW294** | AGAGTTTTGGATGGCTGGCT | **Human NCL qPCR** |
| **SW297** | GCCATCTCGGGCTTTGTCT | **Human UBTF qPCR** |
| **SW298** | CAAAGCCACCTCACCCGAG | **Human UBTF qPCR** |
| **RB CTDSPL F** | TCTTCAAAGGATGGGAGAGC | **CTDSPL qPCR** |
| **RB CTDSPL R** | AGCCGACTCAGATCCTTCAC | **CTDSPL qPCR** |

| | | |
|---|---|---|
| **SW243** | CCCCGCGAACTACCTTACTG | **CTDSPL2 qPCR** |
| **SW244** | CAGCCTCAACAGCTTGTCCT | **CTDSPL2 qPCR** |
| **SW74** | GTAAGACTAAGCCGTGTTGTTG | **chTERT qPCR ex5** |
| **SW75** | CTCCGAATACTGAAGAGC | **chTERT qPCR ex6** |
| **SW70** | AACATGAAATGCAAATTGACTGC | **chTERT qPCR ex11** |
| **SW71** | ACTGTCTGAAGGCTGTTGATCT | **chTERT qPCR ex12** |
| **SW62** | CAGACTACTTTACCTCTTGACACAG | **chTAPAS qPCR ex4** |
| **SW63** | ATGGTGAGCCTTGTGTTGGC | **chTAPAS qPCR ex5** |
| **SM50** | GTGCTGCAGCTCCCATTTCAT | **hTERT qPCR** |
| **SM51** | GCTTTCAGGATGGAGTAGCAGA | **hTERT qPCR** |
| **SW150** | TGTAGCTGAGGTCGGCAAAC | **hTAPAS qPCR** |
| **SW151** | GGTGCGAGGCCTGTTCAAAT | **hTAPAS qPCR** |
| **JJ227** | TGCCATCACAGCCACACAGAAG | **GAPDH qPCR** |
| **JJ228** | ACTTTCCCCACAGCCTTAGCAG | **GAPDH qPCR** |

**References**

Abe, T., Sugimura, K., Hosono, Y., Takami, Y., Akita, M., Yoshimura, A., Tada, S., Nakayama, T., Murofushi, H., Okumura, K., et al. (2011). The histone chaperone facilitates chromatin transcription (FACT) protein maintains normal replication fork rates. J. Biol. Chem. *286*, 30504–30512.

Abramoff, M., Magalhaes, P., and Ram, S. Image Processing with ImageJ. Biophotonics Int.

Aiyer, S., Swapna, G.V.T., Malani, N., Aramini, J.M., Schneider, W.M., Plumb, M.R., Ghanem, M., Larue, R.C., Sharma, A., Studamire, B., et al. (2014). Altering murine leukemia virus integration through disruption of the integrase and BET protein family interaction. Nucleic Acids Res. *42*, 5917–5928.

Angelov, D., Bondarenko, V.A., Almagro, S., Menoni, H., Mongélard, F., Hans, F., Mietton, F., Studitsky, V.M., Hamiche, A., Dimitrov, S., et al. (2006). Nucleolin is a histone chaperone with FACT-like activity and assists remodeling of nucleosomes. EMBO J. *25*, 1669–1679.

El Ashkar, S., De Rijck, J., Demeulemeester, J., Vets, S., Madlala, P., Cermakova, K., Debyser, Z., and Gijsbers, R. (2014). BET-independent MLV-based Vectors Target Away From Promoters and Regulatory Elements. Mol. Ther. Nucleic Acids *3*, e179.

Baba, T.W., and Humphries, E.H. (1986). Selective integration of avian leukosis virus in different hematopoietic tissues. Virology *155*, 557–566.

Ballandras-Colas, A., Brown, M., Cook, N.J., Dewdney, T.G., Demeler, B., Cherepanov,

P., Lyumkis, D., and Engelman, A.N. (2016). Cryo-EM reveals a novel octameric integrase structure for betaretroviral intasome function. Nature *530*, 358–361.

Barnard, R.J.O., and Young, J.A.T. (2003). Alpharetrovirus Envelope-Receptor Interactions. In Cellular Factors Involved in Early Steps of Retroviral Replication, pp. 107–136.

Barr, S.D., Leipzig, J., Shinn, P., Ecker, J.R., and Bushman, F.D. (2005). Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. J. Virol. *79*, 12035–12044.

Beemon, K., and Rosenberg, N. (2012). Mechanisms of oncogenesis by avian and murine retroviruses. In Cancer Associated Viruses, Robertson, ES., ed. (New York, NY: Springer), pp. 677–704.

Belotserkovskaya, R., and Reinberg, D. (2004). Facts about FACT and transcript elongation through chromatin. Curr. Opin. Genet. Dev. *14*, 139–146.

Benleulmi, M., Matysiak, J., Henriquez, D., Vaillant, C., Lesbats, P., Calmels, C., Naughtin, M., Leon, O., Skalka, A., Ruff, M., et al. (2015). Intasome architecture and chromatin density modulate retroviral integration into nucleosome. Retrovirology *12*, 13.

Bisgrove, D., Mahmoudi, T., Henklein, P., and Verdin, E. (2007). Conserved P-TEFb-interacting domain of BRD4 inhibits HIV transcription. Proc. Natl. Acad. Sci. U. S. A. *104*, 13690–13695.

Blasco, M.A. (2005). Telomeres and human disease: ageing, cancer and beyond. Nat. Rev. Genet. *6*, 611–622.

Bodnar, A.G., Ouellette, M., Frolkis, M., Holt, S.E., Chiu, C.-P., Morin, G.B., Harley, C.B., Shay, J.W., Lichtsteiner, S., and Wright, W.E. (1998). Extension of Life-Span by Introduction of Telomerase into Normal Human Cells. Science. *279*.

Bor, Y.C., Bushman, F.D., and Orgel, L.E. (1995). In vitro integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. Proc. Natl. Acad. Sci. U. S. A. *92*, 10334–10338.

Borah, S., Xi, L., Zaug, A.J., Powell, N.M., Dancik, G.M., Cohen, S.B., Costello, J.C., Theodorescu, D., and Cech, T.R. (2015). TERT promoter mutations and telomerase reactivation in urothelial cancer. Science. *347*, 1006–1010.

Bose, S., Basu, M., and Banerjee, A.K. (2004). Role of nucleolin in human parainfluenza virus type 3 infection of human lung epithelial cells. J. Virol. *78*, 8146–8158.

Bowerman, B., Brown, P.O., Bishop, J.M., and Varmus, H.E. (1989). A nucleoprotein complex mediates the integration of retroviral DNA. Genes Dev. *3*, 469–478.

Braconi, C., Valeri, N., Kogure, T., Gasparini, P., Huang, N., Nuovo, G.J., Terracciano, L., Croce, C.M., and Patel, T. (2010). Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. Proc. Natl. Acad. Sci. *108*, 786–791.

Bridier-Nahmias, A., Tchalikian-Cosson, A., Baller, J.A., Menouni, R., Fayol, H., Flores, A., Saib, A., Werner, M., Voytas, D.F., and Lesage, P. (2015). An RNA polymerase III subunit determines sites of retrotransposon integration. Science. *348*, 585–588.

Brown, P.O., Bowerman, B., Varmus, H.E., and Bishop, J.M. (1987). Correct integration

of retroviral DNA in vitro. Cell *49*, 347–356.

Bukrinsky, M., Sharova, N., and Stevenson, M. (1993). Human immunodeficiency virus type 1 2-LTR circles reside in a nucleoprotein complex which is different from the preintegration complex. J. Virol. *67*, 6863–6865.

Burns, J.C., Friedmann, T., Driever, W., Burrascano, M., and Yee, J.K. (1993). Vesicular stomatitis virus G glycoprotein pseudotyped retroviral vectors: concentration to very high titer and efficient gene transfer into mammalian and nonmammalian cells. Proc. Natl. Acad. Sci. U. S. A. *90*, 8033–8037.

Bushman, F.D., Fujiwara, T., and Craigie, R. (1990). Retroviral DNA integration directed by HIV integration protein in vitro. Science *249*, 1555–1558.

Butler, S.L., Hansen, M.S., and Bushman, F.D. (2001). A quantitative assay for HIV DNA integration in vivo. Nat. Med. *7*, 631–634.

Campisi, J. (2005). Senescent cells, tumor suppression, and organismal aging: Good citizens, bad neighbors. Cell *120*, 513–522.

Cavazzana-Calvo, M., Hacein-Bey, S., de Saint Basile, G., Gross, F., Yvon, E., Nusbaum, P., Selz, F., Hue, C., Certain, S., Casanova, J.L., et al. (2000). Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. Science *288*, 669–672.

Cen, S., Javanbakht, H., Kim, S., Shiba, K., Craven, R., Rein, A., Ewalt, K., Schimmel, P., Musier-Forsyth, K., and Kleiman, L. (2002). Retrovirus-specific packaging of aminoacyl-tRNA synthetases with cognate primer tRNAs. J. Virol. *76*, 13111–13115.

Cepko, C. (2001). Large-Scale Preparation and Concentration of Retrovirus Stocks. In Current Protocols in Molecular Biology, (Hoboken, NJ, USA: John Wiley & Sons, Inc.), pp. 9.12.1–9.12.6.

Cereseto, A., Manganaro, L., Gutierrez, M.I., Terreni, M., Fittipaldi, A., Lusic, M., Marcello, A., and Giacca, M. (2005). Acetylation of HIV-1 integrase by p300 regulates viral integration. EMBO J. *24*, 3070–3081.

Charmetant, J., Moreau, K., Gallay, K., Ballandras, A., Gouet, P., and Ronfort, C. (2011). Functional analyses of mutants of the central core domain of an Avian Sarcoma/Leukemia Virus integrase. Virology *421*, 42–50.

Chen, H., and Engelman, A. (1998). The barrier-to-autointegration protein is a host factor for HIV type 1 integration. Proc. Natl. Acad. Sci. U. S. A. *95*, 15270–15274.

Chen, Y.-L., Liu, C.-D., Cheng, C.-P., Zhao, B., Hsu, H.-J., Shen, C.-L., Chiu, S.-J., Kieff, E., and Peng, C. (2014). Nucleolin is important for Epstein-Barr virus nuclear antigen 1-mediated episome binding, maintenance, and transcription. Proc. Natl. Acad. Sci. *111*, 243–248.

Cherepanov, P., Maertens, G., Proost, P., Devreese, B., Van Beeumen, J., Engelborghs, Y., De Clercq, E., and Debyser, Z. (2003). HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. J. Biol. Chem. *278*, 372–381.

Cherepanov, P., Ambrosio, A.L.B., Rahman, S., Ellenberger, T., and Engelman, A. (2005). Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. Proc. Natl. Acad. Sci. U. S. A. *102*, 17308–17313.

Cherepanov, P., Maertens, G.N., and Hare, S. (2011). Structural insights into the retroviral DNA integration apparatus. Curr. Opin. Struct. Biol. *21*, 249–256.

Ciuffi, A., Llano, M., Poeschla, E., Hoffmann, C., Leipzig, J., Shinn, P., Ecker, J.R., and Bushman, F. (2005). A role for LEDGF/p75 in targeting HIV DNA integration. Nat. Med. *11*, 1287–1289.

Clurman, B.E., and Hayward, W.S. (1989). Multiple proto-oncogene activations in avian leukosis virus-induced lymphomas: evidence for stage-specific events. Mol. Cell. Biol. *9*, 2657–2664.

Coffin, J.M., Hughes, S.H., and Varmus, H.E. (1997). Retroviruses. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Collins, K., and Mitchell, J.R. (2002). Telomerase in the human organism. Oncogene *21*, 564–579.

Core, L.J., and Lis, J.T. (2008). Transcription Regulation Through Promoter-Proximal Pausing of RNA Polymerase II. Science. *319*, 1791–1792.

Craigie, R., and Bushman, F.D. (2014). Host Factors in Retroviral Integration and the Selection of Integration Target Sites. Microbiol. Spectr. *2*.

Crowe, B.L., Larue, R.C., Yuan, C., Hess, S., Kvaratskhelia, M., and Foster, M.P. (2016). Structure of the Brd4 ET domain bound to a C-terminal motif from γ-retroviral integrases reveals a conserved mechanism of interaction. Proc. Natl. Acad. Sci. U. S. A. *113*, 2086–2091.

Dai, J., Xie, W., Brady, T.L., Gao, J., and Voytas, D.F. (2007). Phosphorylation Regulates Integration of the Yeast Ty5 Retrotransposon into Heterochromatin.

Mol. Cell *27*, 289–299.

Darlix, J.L., de Rocquigny, H., Mauffret, O., and Mély, Y. (2014). Retrospective on the all-in-one retroviral nucleocapsid protein. Virus Res. *193*, 2–15.

Debyser, Z., Christ, F., De Rijck, J., and Gijsbers, R. (2015). Host factors for retroviral integration site selection. Trends Biochem. Sci. *40*, 108–116.

Delany, M.E., and Daniels, L.M. (2004). The chicken telomerase reverse transcriptase (chTERT): molecular and cytogenetic characterization with a comparative analysis. Gene *339*, 61–69.

Delany, M.E., Krupkin, A.B., and Miller, M.M. (2000). Organization of telomere sequences in birds: evidence for arrays of extreme length and for in vivo shortening. Cytogenet. Cell Genet. *90*, 139–145.

Demeulemeester, J., De Rijck, J., Gijsbers, R., and Debyser, Z. (2015). Retroviral integration: Site matters. BioEssays. 37 (11), 1202-1214.

Derse, D., Crise, B., Li, Y., Princler, G., Lum, N., Stewart, C., McGrath, C.F., Hughes, S.H., Munroe, D.J., and Wu, X. (2007). Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. J. Virol. *81*, 6731–6741.

Dinman, J.D., Richter, S., Plant, E.P., Taylor, R.C., Hammell, A.B., and Rana, T.M. (2002). The frameshift signal of HIV-1 involves a potential intramolecular triplex RNA structure. Proc. Natl. Acad. Sci. U. S. A. *99*, 5331–5336.

Donehower, L.A., and Varmus, H.E. (1984). A mutant murine leukemia virus with a single missense codon in pol is defective in a function affecting integration. Proc.

Natl. Acad. Sci. U. S. A. *81*, 6461–6465.

Dornburg, R. (2003). The history and principles of retroviral vectors. Front. Biosci. *8*, d818–d835.

Eidahl, J.O., Crowe, B.L., North, J.A., McKee, C.J., Shkriabai, N., Feng, L., Plumb, M., Graham, R.L., Gorelick, R.J., Hess, S., et al. (2013). Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes. Nucleic Acids Res. *41*, 3924–3936.

Engelman, a (1999). In vivo analysis of retroviral integrase structure and function. Adv. Virus Res. *52*, 411–426.

Engelman, A. (1994). Most of the avian genome appears available for retroviral DNA integration. BioEssays *16*, 797–799.

Engelman, A.N., and Cherepanov, P. (2017). Retroviral intasomes arising. Curr. Opin. Struct. Biol. *47*, 23–29.

Engelman, A., Mizuuchi, K., and Craigie, R. (1991). HIV-1 DNA integration: Mechanism of viral DNA cleavage and DNA strand transfer. Cell *67*, 1211–1221.

Farnet, C.M., and Bushman, F.D. (1997). HIV-1 cDNA integration: Requirement of HMG I(Y) protein for function of preintegration complexes in vitro. Cell *88*, 483–492.

Faschinger, A., Rouault, F., Sollner, J., Lukas, A., Salmons, B., Günzburg, W.H., and Indik, S. (2008). Mouse mammary tumor virus integration site selection in human and mouse genomes. J. Virol. *82*, 1360–1367.

Fassati, A. (2006). HIV infection of non-dividing cells: a divisive problem. Retrovirology

*3*, 74.

Feng, Y.X., Copeland, T.D., Henderson, L.E., Gorelick, R.J., Bosche, W.J., Levin, J.G., and Rein, A. (1996). HIV-1 nucleocapsid protein induces “maturation” of dimeric retroviral RNA in vitro. Proc. Natl. Acad. Sci. U. S. A. *93*, 7577–7581.

Ferber, M.J., Montoya, D.P., Yu, C., Aderca, I., McGee, A., Thorland, E.C., Nagorney, D.M., Gostout, B.S., Burgart, L.J., Boix, L., et al. (2003). Integrations of the hepatitis B virus (HBV) and human papillomavirus (HPV) into the human telomerase reverse transcriptase (hTERT) gene in liver and cervical cancers. Oncogene *22*, 3813–3820.

Filippakopoulos, P., Qi, J., Picaud, S., Shen, Y., Smith, W.B., Fedorov, O., Morse, E.M., Keates, T., Hickman, T.T., Felletar, I., et al. (2010). Selective inhibition of BET bromodomains. Nature *468*, 1067–1073.

Firouzi, S., López, Y., Suzuki, Y., Nakai, K., Sugano, S., Yamochi, T., and Watanabe, T. (2014). Development and validation of a new high-throughput method to investigate the clonality of HTLV-1-infected cells based on provirus integration sites. Genome Med. *6*, 46.

Florence, B., and Faller, D. V (2001). You bet-cha: a novel family of transcriptional regulators. Front. Biosci. *6*, D1008–D1018.

Formosa, T. (2012). The role of FACT in making and breaking nucleosomes. Biochim. Biophys. Acta *1819*, 247–255.

Gai, X., and Voytas, D.F. (1998). A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin. Mol. Cell *1*, 1051–

1055.

Gallay, P., Swingler, S., Aiken, C., and Trono, D. (1995). HIV-1 infection of nondividing cells: C-terminal tyrosine phosphorylation of the viral matrix protein is a key regulator. Cell *80*, 379–388.

Gallay, P.A., Hope, T.J., Chin, D., and Trono, D. (1997). HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway. Proc. Natl. Acad. Sci. U. S. A. *94*, 9825–9830.

Gao, G., and Goff, S.P. (1998). Replication defect of moloney murine leukemia virus with a mutant reverse transcriptase that can incorporate ribonucleotides and deoxyribonucleotides. J. Virol. *72*, 5905–5911.

van Gent, D.C., Vink, C., Groeneger, A.A., and Plasterk, R.H. (1993). Complementation between HIV integrase proteins mutated in different domains. EMBO J. *12*, 3261–3267.

Gibb, E.A., Brown, C.J., and Lam, W.L. (2011). The functional role of long non-coding RNA in human carcinomas. Mol. Cancer *10*, 38.

Ginisty, H., Amalric, F., and Bouvet, P. (1998). Nucleolin functions in the first step of ribosomal RNA processing. EMBO J. *17*, 1476–1486.

Goff, S.P. (1990). Retroviral reverse transcriptase: synthesis, structure, and function. J. Acquir. Immune Defic. Syndr. *3*, 817–831.

Goodenow, M.M., and Hayward, W.S. (1987). 5’ Long Terminal Repeats of myc-Associated Proviruses Appear Structurally Intact but Are Functionally Impaired in Tumors Induced by Avian Leukosis Viruses. J. Virol. 2489–2498.

Gowda, S., Rao, A.S., Kim, Y.W., and Guntaka, R. V (1988). Identification of sequences in the long terminal repeat of avian sarcoma virus required for efficient transcription. Virology *162*, 243–247.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J.L., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071–1076.

Gupta, S.S., Maetzig, T., Maertens, G.N., Sharif, A., Rothe, M., Weidner-Glunde, M., Galla, M., Schambach, A., Cherepanov, P., and Schulz, T.F. (2013). Bromo- and Extraterminal Domain Chromatin Regulators Serve as Cofactors for Murine Leukemia Virus Integration. J. Virol. *87*, 12721–12736.

Gutschner, T., Hämmerle, M., and Diederichs, S. (2013). MALAT1 -- a paradigm for long noncoding RNA function in cancer. J. Mol. Med. *91*, 791–801.

Hacein-Bey-Abina, S., Le Deist, F., Carlier, F., Bouneaud, C., Hue, C., De Villartay, J.-P., Thrasher, A.J., Wulffraat, N., Sorensen, R., Dupuis-Girod, S., et al. (2002). Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. N. Engl. J. Med. *346*, 1185–1193.

Hacein-Bey-Abina, S., Garrigue, A., Wang, G.P., Soulier, J., Lim, A., Morillon, E., Clappier, E., Caccavelli, L., Delabesse, E., Beldjord, K., et al. (2008). Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J.

Clin. Invest. *118*, 3132–3142.

Hackett, J.A., and Greider, C.W. (2002). Balancing instability: dual roles for telomerase and telomere dysfunction in tumorigenesis. Oncogene *21*, 619–626.

Haferkamp, S., Tran, S.L., Becker, T.M., Scurr, L.L., Kefford, R.F., and Rizos, H. (2009). The relative contributions of the p53 and pRb pathways in oncogene-induced melanocyte senescence. Aging. *1*, 542–556.

Hamard-Peron, E., and Muriaux, D. (2011). Retroviral matrix and lipids, the intimate interaction. Retrovirology *8*, 15.

Hare, S., Gupta, S.S., Valkov, E., Engelman, A., and Cherepanov, P. (2010). Retroviral intasome assembly and inhibition of DNA strand transfer. Nature *464*, 232–236.

Hatziioannou, T., and Goff, S.P. (2001). Infection of nondividing cells by Rous sarcoma virus. J. Virol. *75*, 9526–9531.

Hayward, W.S., Neel, B.G., and Astrin, S.M. (1981). Activation of a cellular onc gene by promoter insertion in ALV-induced lymphoid leukosis. Nature *290*, 475–480.

Heidenreich, B., Rachakonda, P.S., Hemminki, K., and Kumar, R. (2014). TERT promoter mutations in cancer development. Curr. Opin. Genet. Dev. *24*, 30–37.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell *38*, 576–589.

Heinzinger, N.K., Bukinsky, M.I., Haggerty, S. a, Ragland, a M., Kewalramani, V., Lee, M. a, Gendelman, H.E., Ratner, L., Stevenson, M., and Emerman, M. (1994). The

Vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells. Proc. Natl. Acad. Sci. U. S. A. *91*, 7311–7315.

Hirano, M., Kaneko, S., Yamashita, T., Luo, H., Qin, W., Shirota, Y., Nomura, T., Kobayashi, K., and Murakami, S. (2003). Direct interaction between nucleolin and hepatitis C virus NS5B. J. Biol. Chem. *278*, 5109–5115.

Hishinuma, F., DeBona, P.J., Astrin, S., and Skalka, A.M. (1981). Nucleotide sequence of acceptor site and termini of integrated avian endogenous provirus ev1: Integration creates a 6 bp repeat of host DNA. Cell *23*, 155–164.

Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. Science *339*, 959–961.

Howe, S.J., Mansour, M.R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., Brugman, M.H., Pike-Overzet, K., Chatters, S.J., de Ridder, D., et al. (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J. Clin. Invest. *118*, 3143–3150.

Hrdlicková, R., Nehyba, J., and Bose, H.R. (2012). Alternatively spliced telomerase reverse transcriptase variants lacking telomerase activity stimulate cell proliferation. Mol. Cell. Biol. *32*, 4283–4296.

Huang, D.W., Lempicki, R. a, and Sherman, B.T. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*,

44–57.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G. V, Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science. *339*, 957–959.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., et al. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell *142*, 409–419.

Hughes, S.H. (2004). The RCAS vector system. Folia Biol. (Praha). *50*, 107–119.

Ikeda, Y., Kinoshita, Y., Susaki, D., Ikeda, Y., Iwano, M., Takayama, S., Higashiyama, T., Kakutani, T., and Kinoshita, T. (2011). HMG domain containing SSRP1 is required for DNA demethylation and genomic imprinting in Arabidopsis. Dev. Cell *21*, 589–596.

Ilves, I., Mäemets, K., Silla, T., Janikson, K., and Ustav, M. (2006). Brd4 is involved in multiple processes of the bovine papillomavirus type 1 life cycle. J. Virol. *80*, 3660–3665.

Jin, D.P., Li, C.Y., Yang, H.J., Zhang, W.X., Li, C.L., Guan, W.J., and Ma, Y.H. (2011). Apoptotic effects of hydrogen peroxide and vitamin C on chicken embryonic fibroblasts: redox state and programmed cell death. Cytotechnology *63*, 461–471.

Justice, J., and Beemon, K.L. (2013). Avian retroviral replication. Curr. Opin. Virol. *3*, 664–669.

Justice, J., Malhotra, S., Ruano, M., Li, Y., Zavala, G., Lee, N., Morgan, R., and Beemon,

K. (2015a). The MET Gene Is a Common Integration Target in Avian Leukosis Virus Subgroup J-Induced Chicken Hemangiomas. J. Virol. *89*, 4712–4719.

Justice, J.F., Morgan, R.W., and Beemon, K.L. (2015b). Common Viral Integration Sites Identified in Avian Leukosis Virus-Induced B-Cell Lymphomas. MBio *6*, e01863–15.

Kalpana, G. V, Marmon, S., Wang, W., Crabtree, G.R., and Goff, S.P. (1994). Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. Science *266*, 2002–2006.

Kanter, M.R., Smith, R.E., and Hayward, W.S. (1988). Rapid induction of B-cell lymphomas: insertional activation of c-myb by avian leukosis virus. J. Virol. *62*, 1423–1432.

Karageorgos, L., Li, P., and Burrell, C.J. (1995). Stepwise analysis of reverse transcription in a cell-to-cell human immunodeficiency virus infection model: kinetics and implications. J Gen Virol *76*, 1675–1686.

Kashuba, V.I., Li, J., Wang, F., Senchenko, V.N., Protopopov, A., Malyukova, A., Kutsenko, A.S., Kadyrova, E., Zabarovska, V.I., Muravenko, O. V, et al. (2004). RBSP3 (HYA22) is a tumor suppressor gene implicated in major epithelial malignancies. Proc. Natl. Acad. Sci. U. S. A. *101*, 4906–4911.

Kashuba, V.I., Pavlova, T. V, Grigorieva, E. V, Kutsenko, A., Yenamandra, S.P., Li, J., Wang, F., Protopopov, A.I., Zabarovska, V.I., Senchenko, V., et al. (2009). High mutability of the tumor suppressor genes RASSF1 and RBSP3 (CTDSPL) in cancer. PLoS One *4*, e5231.

Katz, R.A., Merkel, G., Kulkosky, J., Leis, J., and Skalka, A.M. (1990). The avian retroviral IN protein is both necessary and sufficient for integrative recombination in vitro. Cell *63*, 87–95.

Kessl, J.J., Jena, N., Koh, Y., Taskent-Sezgin, H., Slaughter, A., Feng, L., De Silva, S., Wu, L., Le Grice, S.F.J., Engelman, A., et al. (2012). Multimode, cooperative mechanism of action of allosteric HIV-1 integrase inhibitors. J. Biol. Chem. *287*, 16801–16811.

Kessl, J.J., Kutluay, S.B., Townsend, D., Rebensburg, S., Slaughter, A., Larue, R.C., Shkriabai, N., Bakouche, N., Fuchs, J.R., Bieniasz, P.D., et al. (2016). HIV-1 Integrase Binds the Viral RNA Genome and Is Essential during Virion Morphogenesis. Cell *166*, 1257–1268.e12.

Killela, P.J., Reitman, Z.J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, L.A., Friedman, A.H., Friedman, H., Gallia, G.L., Giovanella, B.C., et al. (2013). TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. Proc. Natl. Acad. Sci. U. S. A. *110*, 6021–6026.

Kirk, P.D.W., Huvet, M., Melamed, A., Maertens, G.N., and Bangham, C.R.M. (2016). Retroviruses integrate into a shared, non-palindromic DNA motif. Nat. Microbiol. *2*, 16212.

Kloet, D.E.A., Polderman, P.E., Eijkelenboom, A., Smits, L.M., van Triest, M.H., van den Berg, M.C.W., Groot Koerkamp, M.J., van Leenen, D., Lijnzaad, P., Holstege, F.C., et al. (2015). FOXO target gene CTDSP2 regulates cell cycle progression through Ras and p21(Cip1/Waf1). Biochem. J. *469*, 289–298.

Knockaert, M., Sapkota, G., Alarcón, C., Massagué, J., and Brivanlou, A.H. (2006). Unique players in the BMP pathway: small C-terminal domain phosphatases dephosphorylate Smad1 to attenuate BMP signaling. Proc. Natl. Acad. Sci. U. S. A. *103*, 11940–11945.

Koh, C.M., Khattar, E., Leow, S.C., Liu, C.Y., Muller, J., Ang, W.X., Li, Y., Franzoso, G., Li, S., Guccione, E., et al. (2015). Telomerase regulates MYC-driven oncogenesis independent of its reverse transcriptase activity. J. Clin. Invest. *125*, 2109–2122.

Krishnan, L., and Engelman, A. (2012). Retroviral integrase proteins and HIV-1 DNA integration. J. Biol. Chem. *287*, 40858–40866.

Kumari, A., Mazina, O.M., Shinde, U., Mazin, A. V, and Lu, H. (2009). A role for SSRP1 in recombination-mediated DNA damage response. J. Cell. Biochem. *108*, 508–518.

Kvaratskhelia, M., Sharma, A., Larue, R.C., Serrao, E., and Engelman, A. (2014). Molecular mechanisms of retroviral integration site selection. Nucleic Acids Res. *42*, 10209–10225.

Kwon, H., and Green, M.R. (1994). The RNA polymerase I transcription factor, upstream binding factor, interacts directly with the TATA box-binding protein. J. Biol. Chem. *269*, 30140–30146.

Lafave, M.C., Varshney, G.K., Gildea, D.E., Wolfsberg, T.G., Baxevanis, A.D., and Burgess, S.M. (2014). MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Res. *42*, 4257–4269.

Laimins, L.A., Tsichlis, P., and Khoury, G. (1984). Multiple enhancer domains in the 3' terminus of the Prague strain of Rous sarcoma virus. Nucleic Acids Res. *12*, 6427–6442.

Larue, R.C., Plumb, M.R., Crowe, B.L., Shkriabai, N., Sharma, A., DiFiore, J., Malani, N., Aiyer, S.S., Roth, M.J., Bushman, F.D., et al. (2014). Bimodal high-affinity association of Brd4 with murine leukemia virus integrase and mononucleosomes. Nucleic Acids Res. *42*, 4868–4881.

Lee, M.S., and Craigie, R. (1994). Protection of retroviral DNA from autointegration: involvement of a cellular factor. Proc. Natl. Acad. Sci. U. S. A. *91*, 9823–9827.

Lee, M.S., and Craigie, R. (1998). A previously unidentified host protein protects retroviral DNA from autointegration. Proc. Natl. Acad. Sci. U. S. A. *95*, 1528–1533.

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The sequence read archive. Nucleic Acids Res. *39*, D19–D21.

LeRoy, G., Rickards, B., and Flint, S.J. (2008). The Double Bromodomain Proteins Brd2 and Brd3 Couple Histone Acetylation to Transcription. Mol. Cell *30*, 51–60.

Lesage, P., and Todeschini, A.L. (2005). Happy together: The life and times of Ty retrotransposons and their hosts. Cytogenet. Genome Res. *110*, 70–90.

Lewinski, M.K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., Collins, F., Shinn, P., Leipzig, J., Hannenhalli, S., et al. (2006). Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS Pathog. *2*, e60.

Li, L., Olvera, J.M., Yoder, K.E., Mitchell, R.S., Butler, S.L., Lieber, M., Martin, S.L., and Bushman, F.D. (2001). Role of the non-homologous DNA end joining pathway in the early steps of retroviral infection. EMBO J. *20*, 3272–3281.

Li, Y., Liu, X., Yang, Z., Xu, C., Liu, D., Qin, J., Dai, M., Hao, J., Feng, M., Huang, X., et al. (2014). The MYC, TERT, and ZIC1 genes are common targets of viral integration and transcriptional deregulation in avian leukosis virus subgroup J-induced myeloid leukosis. J. Virol. *88*, 3182–3191.

Liang, C.-C., Park, A.Y., and Guan, J.-L. (2007). In vitro scratch assay: a convenient and inexpensive method for analysis of cell migration in vitro. Nat. Protoc. *2*, 329–333.

Lin, A., Wang, S., Nguyen, T., Shire, K., and Frappier, L. (2008). The EBNA1 protein of Epstein-Barr virus functionally interacts with Brd4. J. Virol. *82*, 12009–12019.

Llano, M., Saenz, D.T., Meehan, A., Wongthida, P., Peretz, M., Walker, W.H., Teo, W., and Poeschla, E.M. (2006). An essential role for LEDGF/p75 in HIV integration. Science. *314*, 461–464.

Lu, K., Heng, X., and Summers, M.F. (2011). Structural determinants and mechanism of HIV-1 genome packaging. J. Mol. Biol. *410*, 609–633.

Lu, R., Ghory, H.Z., and Engelman, A. (2005). Genetic Analyses of Conserved Residues in the Carboxyl-Terminal Domain of Human Immunodeficiency Virus Type 1 Integrase. J. Virol. *79*, 10356–10368.

Maegdefrau, U., and Bosserhoff, A.-K. (2012). BMP activated Smad signaling strongly promotes migration and invasion of hepatocellular carcinoma cells. Exp. Mol.

Pathol. *92*, 74–81.

Maertens, G.N. (2016). B'-protein phosphatase 2A is a functional binding partner of delta-retroviral integrase. Nucleic Acids Res. *44*, 364–376.

Maertens, G., Cherepanov, P., Pluymers, W., Busschots, K., De Clercq, E., Debyser, Z., and Engelborghs, Y. (2003). LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. J. Biol. Chem. *278*, 33528–33539.

Maertens, G.N., Hare, S., and Cherepanov, P. (2010). The mechanism of retroviral integration from X-ray structures of its key intermediates. Nature *468*, 326–329.

Malhotra, S., Winans, S., Lam, G., Justice, J., Morgan, R., and Beemon, K. (2017). Selection for avian leukosis virus integration sites determines the clonal progression of B-cell lymphomas. PLOS Pathog. *13*, e1006708.

Mandal, D., and Prasad, V.R. (2009). Analysis of 2-LTR circle junctions of viral DNA in infected cells. Methods Mol. Biol. *485*, 73–85.

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., et al. (2014). CDD: NCBI's conserved domain database. Nucleic Acids Res. *43*, D222–D226.

Matysiak, J., Lesbats, P., Mauro, E., Lapaillerie, D., Dupuy, J.-W., Lopez, A.P., Benleulmi, M.S., Calmels, C., Andreola, M.-L., Ruff, M., et al. (2017). Modulation of chromatin structure by the FACT histone chaperone complex regulates HIV-1 integration. Retrovirology *14*, 39.

McCormack, M.P., and Rabbitts, T.H. (2004). Activation of the T-cell oncogene LMO2

after gene therapy for X-linked severe combined immunodeficiency. N. Engl. J. Med. *350*, 913–922.

McKee, C.J., Kessl, J.J., Shkriabai, N., Dar, M.J., Engelman, A., and Kvaratskhelia, M. (2008). Dynamic modulation of HIV-1 integrase structure and function by cellular lens epithelium-derived growth factor (LEDGF) protein. J. Biol. Chem. *283*, 31802–31812.

Mitchell, R.S., Beitzel, B.F., Schroder, A.R.W., Shinn, P., Chen, H., Berry, C.C., Ecker, J.R., and Bushman, F.D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. *2*, E234.

Mongelard, F., and Bouvet, P. (2007). Nucleolin: a multiFACeTed protein. Trends Cell Biol. *17*, 80–86.

Mourtada-Maarabouni, M., Pickard, M.R., Hedge, V.L., Farzaneh, F., and Williams, G.T. (2009). GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. Oncogene *28*, 195–208.

Müller, H.P., and Varmus, H.E. (1994). DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. EMBO J. *13*, 4704–4714.

Narezkina, A., Taganov, K.D., Litwin, S., Stoyanova, R., Hayashi, J., Seeger, C., Skalka, A.M., and Katz, R.A. (2004). Genome-wide analyses of avian sarcoma virus integration sites. J. Virol. *78*, 11656–11663.

Neel, B.G., Hayward, W.S., Robinson, H.L., Fang, J., and Astrin, S.M. (1981). Avian leukosis virus-induced tumors have common proviral integration sites and

synthesize discrete new RNAs: oncogenesis by promoter insertion. Cell *23*, 323–334.

Nowotny, M. (2009). Retroviral integrase superfamily: the structural perspective. EMBO Rep. *10*, 144–151.

O'Sullivan, C.T., Polony, T.S., Paca, R.E., and Beemon, K.L. (2002). Rous Sarcoma Virus Negative Regulator of Splicing Selectively Suppresses src mRNA Splicing and Promotes Polyadenylation. Virology *302*, 405–412.

Okada, M., Okawa, K., Isobe, T., and Fukagawa, T. (2009). CENP-H-containing complex facilitates centromere deposition of CENP-A in cooperation with FACT and CHD1. Mol. Biol. Cell *20*, 3986–3995.

Oliveira, D. V, Kato, A., Nakamura, K., Ikura, T., Okada, M., Kobayashi, J., Yanagihara, H., Saito, Y., Tauchi, H., and Komatsu, K. (2014). Histone chaperone FACT regulates homologous recombination by chromatin remodeling through interaction with RNF20. J. Cell Sci. *127*, 763–772.

Orphanides, G., LeRoy, G., Chang, C.H., Luse, D.S., and Reinberg, D. (1998). FACT, a factor that facilitates transcript elongation through nucleosomes. Cell *92*, 105–116.

Orphanides, G., Wu, W.H., Lane, W.S., Hampsey, M., and Reinberg, D. (1999). The chromatin-specific transcription elongation factor FACT comprises human SPT16 and SSRP1 proteins. Nature *400*, 284–288.

Ouellet Lavallée, G., and Pearson, A. (2015). Upstream binding factor inhibits herpes simplex virus replication. Virology *483*, 108–116.

Owens, P., Pickup, M.W., Novitskiy, S. V, Giltnane, J.M., Gorska, A.E., Hopkins, C.R., Hong, C.C., and Moses, H.L. (2015). Inhibition of BMP signaling suppresses metastasis in mammary cancer. Oncogene *34*, 2437–2449.

Panganiban, A.T., and Temin, H.M. (1984). Circles with two tandem LTRs are precursors to integrated retrovirus DNA. Cell *36*, 673–679.

Polager, S., and Ginsberg, D. (2008). E2F - at the crossroads of life and death. Trends Cell Biol. *18*, 528–535.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., and Pandolfi, P.P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature *465*, 1033–1038.

Polony, T.S., Bowers, S.J., Neiman, P.E., and Beemon, K.L. (2003). Silent point mutation in an avian retrovirus RNA processing element promotes c-myb-associated short-latency lymphomas. J. Virol. *77*, 9378–9387.

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., et al. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat. Biotechnol. *29*, 742–749.

Pryciak, P.M., and Varmus, H.E. (1992). Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell *69*, 769–780.

Qiu, J., and Brown, K.E. (1999). A 110-kDa nuclear shuttle protein, nucleolin, specifically binds to adeno-associated virus type 2 (AAV-2) capsid. Virology *257*,

373–382.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rans, T.S., and England, R. (2009). The evolution of gene therapy in X-linked severe combined immunodeficiency. Ann. Allergy. Asthma Immunol. *102*, 357–362; 363–365, 402.

De Ravin, S.S., Su, L., Theobald, N., Choi, U., Macpherson, J.L., Poidinger, M., Symonds, G., Pond, S.M., Ferris, A.L., Hughes, S.H., et al. (2014). Enhancers are major targets for murine leukemia virus vector integration. J. Virol. *88*, 4504–4513.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. *35*, W193–W200.

Reinberg, D., and Sims, R.J. (2006). de FACTo nucleosome dynamics. J. Biol. Chem. *281*, 23297–23301.

Rickards, B., Flint, S.J., Cole, M.D., and LeRoy, G. (2007). Nucleolin is required for RNA polymerase I transcription in vivo. Mol. Cell. Biol. *27*, 937–948.

De Rijck, J., de Kogel, C., Demeulemeester, J., Vets, S., El Ashkar, S., Malani, N., Bushman, F.D., Landuyt, B., Husson, S.J., Busschots, K., et al. (2013). The BET family of proteins targets moloney murine leukemia virus integration near transcription start sites. Cell Rep. *5*, 886–894.

Rous, P. (1910). A transmissible avian neoplasm. (Sarcoma of the common fowl.). J.

Exp. Med. *12*, 696–705.

Saebøe-Larssen, S., Fossberg, E., and Gaudernack, G. (2006). Characterization of novel alternative splicing sites in human telomerase reverse transcriptase (hTERT): analysis of expression and mutual correlation in mRNA isoforms from normal and tumour tissues. BMC Mol. Biol. *7*, 26.

Safina, A., Garcia, H., Commane, M., Guryanova, O., Degan, S., Kolesnikova, K., and Gurova, K. V (2013). Complex mutual regulation of facilitates chromatin transcription (FACT) subunits on both mRNA and protein levels in human cells. Cell Cycle *12*, 2423–2434.

Sanij, E., Diesch, J., Lesmana, A., Poortinga, G., Hein, N., Lidgerwood, G., Cameron, D.P., Ellul, J., Goodall, G.J., Wong, L.H., et al. (2015). A novel role for the pol I transcription factor ubtf in maintaining genome stability through the regulation of highly transcribed pol II genes. Genome Res. *25*, 201–212.

Schröder, A.R.W., Shinn, P., Chen, H., Berry, C., Ecker, J.R., and Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. Cell *110*, 521–529.

Senchenko, V.N., Anedchenko, E.A., Kondratieva, T.T., Krasnov, G.S., Dmitriev, A.A., Zabarovska, V.I., Pavlova, T. V, Kashuba, V.I., Lerman, M.I., and Zabarovsky, E.R. (2010). Simultaneous down-regulation of tumor suppressor genes RBSP3/CTDSPL, NPRL2/G21 and RASSF1A in primary non-small cell lung cancer. BMC Cancer *10*, 75.

Sharma, A., Larue, R.C., Plumb, M.R., Malani, N., Male, F., Slaughter, A., Kessl, J.J.,

Shkriabai, N., Coward, E., Aiyer, S.S., et al. (2013). BET proteins promote efficient murine leukemia virus integration at transcription start sites. Proc. Natl. Acad. Sci. U. S. A. *110*, 12036–12041.

Shay, J.W., and Wright, W.E. (2011). Role of telomeres and telomerase in cancer. Semin. Cancer Biol. *21*, 349–353.

Shu-Yun Le, Shapiro, B.A., Chen, J.H., Nussinov, R., and Maizel, J. V. (1991). RNA pseudoknots downstream of the frameshift sites of retroviruses. Genet. Anal. Biomol. Eng. *8*, 191–205.

Shun, M.-C., Raghavendra, N.K., Vandegraaff, N., Daigle, J.E., Hughes, S., Kellam, P., Cherepanov, P., and Engelman, A. (2007). LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. Genes Dev. *21*, 1767–1778.

Simon, M.C., Neckameyer, W.S., Hayward, W.S., and Smith, R.E. (1987). Genetic determinants of neoplastic diseases induced by a subgroup F avian leukosis virus. J. Virol. *61*, 1203–1212.

Singh, P.K., Plumb, M.R., Ferris, A.L., Iben, J.R., Wu, X., Fadel, H.J., Luke, B.T., Esnault, C., Poeschla, E.M., Hughes, S.H., et al. (2015). LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. Genes Dev. *29*, 2287–2297.

Smith, M.R., Smith, R.E., Dunkel, I., Hou, V., Beemon, K.L., and Hayward, W.S. (1997). Genetic determinant of rapid-onset B-cell lymphoma by avian leukosis virus. J. Virol. *71*, 6534–6540.

Sowd, G.A., Serrao, E., Wang, H., Wang, W., Fadel, H.J., Poeschla, E.M., and Engelman, A.N. (2016). A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. Proc. Natl. Acad. Sci. U. S. A. *113*, E1054–E1063.

Suerth, J.D., Labenski, V., and Schambach, A. (2014). Alpharetroviral vectors: from a cancer-causing agent to a useful tool for human gene therapy. Viruses *6*, 4811–4838.

Suzuki, Y., and Craigie, R. (2002). Regulatory mechanisms by which barrier-to-autointegration factor blocks autointegration and stimulates intermolecular integration of Moloney murine leukemia virus preintegration complexes. J. Virol. *76*, 12376–12380.

Suzuki, Y., Chew, M.L., and Suzuki, Y. (2012). Role of host-encoded proteins in restriction of retroviral integration. Front. Microbiol. *3*.

Tam, W., Ben-Yehuda, D., and Hayward, W.S. (1997). bic, a novel gene activated by proviral insertions in avian leukosis virus-induced lymphomas, is likely to function through its noncoding RNA. Mol. Cell. Biol. *17*, 1490–1502.

Taylor, H.A., and Delany, M.E. (2000). Ontogeny of telomerase in chicken: Impact of downregulation on pre- and postnatal telomere length in vivo. Dev. Growth Differ. *42*, 613–621.

Tayyari, F., Marchant, D., Moraes, T.J., Duan, W., Mastrangelo, P., and Hegele, R.G. (2011). Identification of nucleolin as a cellular receptor for human respiratory syncytial virus. Nat. Med. *17*, 1132–1135.

Thompson, J., Lepikhova, T., Teixido-Travesa, N., Whitehead, M.A., Palvimo, J.J., and Jänne, O.A. (2006). Small carboxyl-terminal domain phosphatase 2 attenuates androgen-dependent transcription. EMBO J. *25*, 2757–2767.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Trobridge, G.D., Miller, D.G., Jacobs, M.A., Allen, J.M., Kiem, H.-P., Kaul, R., and Russell, D.W. (2006). Foamy virus vector integration sites in normal human cells. Proc. Natl. Acad. Sci. U. S. A. *103*, 1498–1503.

Vink, C., Lutzke, R.A., and Plasterk, R.H. (1994). Formation of a stable complex between the human immunodeficiency virus integrase protein and viral DNA. Nucleic Acids Res. *22*, 4103–4110.

Visvanathan, J., Lee, S., Lee, B., Lee, J.W., and Lee, S.-K. (2007). The microRNA miR-124 antagonizes the anti-neural REST/SCP1 pathway during embryonic CNS development. Genes Dev. *21*, 744–749.

Wakano, C., Byun, J.S., Di, L.J., and Gardner, K. (2012). The dual lives of bidirectional promoters. Biochim. Biophys. Acta - Gene Regul. Mech. *1819*, 688–693.

Wang, W., Liao, P., Shen, M., Chen, T., Chen, Y., Li, Y., Lin, X., Ge, X., and Wang, P. (2016). SCP1 regulates c-Myc stability and functions through dephosphorylating c-Myc Ser62. Oncogene *35*, 491–500.

Wani, S., Sugita, A., Ohkuma, Y., and Hirose, Y. (2016). Human SCP4 is a chromatin-

associated CTD phosphatase and exhibits the dynamic translocation during erythroid differentiation. J. Biochem. *160*, 111–120.

Wei, W., Pelechano, V., Järvelin, A.I., and Steinmetz, L.M. (2011). Functional consequences of bidirectional promoters. Trends Genet. *27*, 267–276.

Werner, S., Hindmarsh, P., Napirei, M., Vogel-Bachmayr, K., and Wöhrl, B.M. (2002). Subcellular localization and integration activities of rous sarcoma virus reverse transcriptase. J. Virol. *76*, 6205–6212.

Winding, P., and Berchtold, M.W. (2001). The chicken B cell line DT40: A novel tool for gene disruption experiments. J. Immunol. Methods *249*, 1–16.

Winkler, D.D., and Luger, K. (2011). The histone chaperone FACT: structural insights and mechanisms for nucleosome reorganization. J. Biol. Chem. *286*, 18369–18374.

Winkler, D.D., Muthurajan, U.M., Hieb, A.R., and Luger, K. (2011). Histone chaperone FACT coordinates nucleosome interaction through multiple synergistic binding events. J. Biol. Chem. *286*, 41883–41892.

Withers, J.B., Ashvetiya, T., and Beemon, K.L. (2012). Exclusion of exon 2 is a common mRNA splice variant of primate telomerase reverse transcriptases. PLoS One *7*, e48016.

Withers-Ward, E.S., Kitamura, Y., Barnes, J.P., and Coffin, J.M. (1994). Distribution of targets for avian retrovirus DNA integration in vivo. Genes Dev. *8*, 1473–1487.

Wrighton, K.H., Willis, D., Long, J., Liu, F., Lin, X., and Feng, X.-H. (2006). Small C-terminal Domain Phosphatases Dephosphorylate the Regulatory Linker Regions

of Smad2 and Smad3 to Enhance Transforming Growth Factor-beta Signaling. J. Biol. Chem. *281*, 38365–38375.

Wu, X., Li, Y., Crise, B., and Burgess, S.M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. Science *300*, 1749–1751.

Wu, Y., Evers, B.M., and Zhou, B.P. (2009). Small C-terminal domain phosphatase enhances snail activity through dephosphorylation. J. Biol. Chem. *284*, 640–648.

Xie, W., Gai, X., Zhu, Y., Zappulla, D.C., Sternglanz, R., and Voytas, D.F. (2001). Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. Mol. Cell. Biol. *21*, 6606–6614.

Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K., and Uversky, V.N. (2010). PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. Biochim. Biophys. Acta - Proteins Proteomics *1804*, 996–1010.

Yamashita, M., and Emerman, M. (2006). Retroviral infection of non-dividing cells: old and new perspectives. Virology *344*, 88–93.

Yang, F., Xian, R.R., Li, Y., Polony, T.S., and Beemon, K.L. (2007a). Telomerase reverse transcriptase expression elevated by avian leukosis virus integration in B cell lymphomas. Proc. Natl. Acad. Sci. U. S. A. *104*, 18952–18957.

Yang, F., Xian, R.R., Li, Y., Polony, T.S., and Beemon, K.L. (2007b). Telomerase reverse transcriptase expression elevated by avian leukosis virus integration in B cell lymphomas. Proc. Natl. Acad. Sci. U. S. A. *104*, 18952–18957.

Yeo, M., Lin, P.S., Dahmus, M.E., and Gill, G.N. (2003). A novel RNA polymerase II C-terminal domain phosphatase that preferentially dephosphorylates serine 5. J.

Biol. Chem. *278*, 26078–26085.

Yeo, M., Lee, S.-K., Lee, B., Ruiz, E.C., Pfaff, S.L., and Gill, G.N. (2005). Small CTD phosphatases function in silencing neuronal gene expression. Science *307*, 596–600.

Yin, Z., Shi, K., Banerjee, S., Pandey, K.K., Bera, S., Grandgenett, D.P., and Aihara, H. (2016). Crystal structure of the Rous sarcoma virus intasome. Nature *530*, 362–366.

You, J., Srinivasan, V., Denis, G. V, Harrington, W.J., Ballestas, M.E., Kaye, K.M., and Howley, P.M. (2006). Kaposi's sarcoma-associated herpesvirus latency-associated nuclear antigen interacts with bromodomain protein Brd4 on host mitotic chromosomes. J. Virol. *80*, 8909–8919.

Zhang, H., Zhou, Y., Alcock, C., Kiefer, T., Monie, D., Siliciano, J., Li, Q., Pham, P., Cofrancesco, J., Persaud, D., et al. (2004). Novel single-cell-level phenotypic assay for residual drug susceptibility and reduced replication capacity of drug-resistant human immunodeficiency virus type 1. J. Virol. *78*, 1718–1729.

Zhang, X., Gejman, R., Mahta, A., Zhong, Y., Rice, K.A., Zhou, Y., Cheunsuchon, P., Louis, D.N., and Klibanski, A. (2010). Maternally expressed gene 3, an imprinted noncoding RNA gene, is associated with meningioma pathogenesis and progression. Cancer Res. *70*, 2350–2358.

Zhao, J., Sun, B.K., Erwin, J.A., Song, J.-J., and Lee, J.T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science *322*, 750–756.

Zhao, Y., Xiao, M., Sun, B., Zhang, Z., Shen, T., Duan, X., Yu, P.B., Feng, X.-H., and Lin, X. (2014). C-terminal domain (CTD) small phosphatase-like 2 modulates the canonical bone morphogenetic protein (BMP) signaling and mesenchymal differentiation via Smad dephosphorylation. J. Biol. Chem. *289*, 26441–26450.

Zheng, Y., and Yao, X. (2013). Posttranslational modifications of HIV-1 integrase by various cellular proteins during viral replication. Viruses *5*, 1787–1801.

Zhu, J., Zhao, Y., and Wang, S. (2010). Chromatin and epigenetic regulation of the telomerase reverse transcriptase gene. Protein Cell *1*, 22–32.

Zhu, J., Gaiha, G.D., John, S.P., Pertel, T., Chin, C.R., Gao, G., Qu, H., Walker, B.D., Elledge, S.J., and Brass, A.L. (2012a). Reactivation of Latent HIV-1 by Inhibition of BRD4. Cell Rep. *2*, 807–816.

Zhu, Y., Lu, Y., Zhang, Q., Liu, J.-J., Li, T.-J., Yang, J.-R., Zeng, C., and Zhuang, S.-M. (2012b). MicroRNA-26a/b and their host genes cooperate to inhibit the G1/S transition by activating the pRb protein. Nucleic Acids Res. *40*, 4615–4625.

Zou, S., Ke, N., Kim, J.M., and Voytas, D.F. (1996). The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. Genes Dev. *10*, 634–645.

**SHELBY WINANS**

swinans1@jhu.edu
(443) 762-9891

## EDUCATION

Johns Hopkins University, Baltimore, MD
Cumulative GPA: 4.0
Ph.D. candidate in Cellular, Molecular, Developmental Biology and Biophysics
Expected graduation date: Dec. 2017
University of Chicago, Chicago, IL
Cumulative GPA: 3.7
Honors B.A. in Biological Sciences; Specialization in Genetics, 2013

## AWARDS & FELLOWSHIPS

Dean's Teaching Fellowship, JHU, Fall 2017
*Retrovirology* Fellowship, 28th International Workshop on Retroviral Pathogenesis, Fall 2016
Victor Corces Teaching Award for Genetics, JHU, Fall 2015
Dean's List, University of Chicago, 2009-2013
Chicago Careers in Health Professions Fellow, University of Chicago, 2011-2013

## RESEARCH EXPERIENCE

**Johns Hopkins University**
*Beemon Laboratory, Baltimore, MD* June 2014 – present
Graduate Researcher
Analysis of avian leukosis virus (ALV) integration and consequences on tumorigenesis.
- Based on ALV insertional mutagenesis, characterized novel genes involved in oncogenesis such as CTDSPL2 and TAPAs RNA (TERT antisense promoter associated RNA).
- Identified host cell factors that regulate ALV proviral integration *in vivo*.

**Undergraduate Honors Thesis Independent Research**
*Dolan Laboratory, University of Chicago, Chicago, IL* March 2012 – June 2013
*Research Assistant*
Identification of "master regulator" SNPs associated with cisplatin sensitivity.
- Genome wide association study to identify SNPs that correlate with cisplatin sensitivity (IC50)..
- Assessed relationship between SNP genotype and cisplatin sensitivity using siRNA to knockdown candidate genes and measure changes in drug sensitivity.

**University of Chicago Medical Center**
*Das Laboratory, University of Chicago, Chicago, IL* June 2011 – June 2013

Genetic testing for hereditary disorders.

- Extraction of DNA, PCR amplification and sequencing of genes for known genetic markers associated with various diseases.

## TEACHING EXPERIENCE

**Johns Hopkins University**
*Baltimore, MD*

| | |
|---|---|
| Instructor, Emerging Infectious Diseases | Fall 2017 |
| Teaching Assistant, Virology | Spring 2016 |
| Teaching Assistant, Eukaryotic Molecular Biology | Fall 2016 |
| Teaching Assistant, Developmental Biology Laboratory | Spring 2015 |
| Teaching Assistant, Genetics Laboratory | Fall 2014 |

## CONFERENCE PRESENTATIONS

**Winans S.,** Larue R., Shkriabai N., Winkler D., Kvaratskhelia M., and Beemon K. Identification of host factors that regulate and target ALV integration. Poster presentation delivered at 6th International Conference on Retroviral Integration, Bordeaux, France, September 2017.

**Winans S.**, Malhotra S., Balagopal V., Li Y., and Beemon K. *CTDSPL2* identified as tumor suppressor gene by ALV insertional mutagenesis. Oral presentation delivered at 28th International Workshop on Retroviral Pathogenesis, New Orleans, LA, November 2016.

**Winans S**., Nehyba J., Malhotra S., Justice J., and Beemon K. ALV integrations into TERT promoter-associated lncRNA in B-cell lymphomas. Poster presentation delivered at 28th International Workshop on Retroviral Pathogenesis, New Orleans, LA, November 2016.

**Winans S**. and Beemon K. The *FACT*s about ALV proviral integration. Oral presentation delivered at the 2016 HIV Dynamics and Replication Program ThinkTank Meeting, Frederick, MD, April 2016.

**Winans S**., Nehyba J., Malhotra S., Justice J., and Beemon K. ALV integrations into TERT promoter-associated lncRNA in B-cell lymphomas. Poster presentation delivered at Cold Spring Harbor Laboratory Retroviruses Meeting, Cold Spring Harbor, NY, May 2015.

## PEER-REVIEWED PUBLICATIONS

**Winans S.**, Larue R., Abraham C., Shkriabai N., Skopp A., Winkler D., Kvaratskhelia M., Beemon K. (2017). The FACT complex promotes avian leukosis virus integration. *J. Virol*. 91(7*).* pii: e00082-17. Spotlight featured article.

**Winans S**., Flynn A., Malhotra S., Balagopal V., Beemon K. (2017). Integration into *CTDSPL* and *CTDSPL2* genes in B-cell lymphomas promotes cell immortalization, migration and survival. *Oncotarget* 8(34):57302-57315.

Nehyba J.*, Malhotra S.*, **Winans S**.*, O'Hare T, Justice J. 4th, Beemon K. (2016). Avian leukosis virus activation of an antisense RNA upstream of *TERT* in B-cell lymphomas. *J Virol*. 90(20):9509-17.
(* indicates co-first authors).

Malhotra S., **Winans S**., Lam G., Justice J., Morgan R.,Beemon K. (2017). Selection for avian leukosis virus integration sites determines the clonal progression of B-cell lymphomas. *PLoS Pathogens* 13(11):e1006708.

Malhotra S., Freeberg M., **Winans S.**, Justice J., Beemon K. (2017). A novel long non-coding RNA in the hTERT promoter region regulates hTERT expression. *ncRNA*. (*manuscript in revision*).

## REFERENCES

**Karen Beemon***,* PhD, Department of Biology, Johns Hopkins University, USA
KLB@jhu.edu, (410) 516-7289

**Evangelos Moudrianakis,** PhD, Department of Biology, Johns Hopkins University, USA
vanm@jhu.edu, (410) 516-7305

**Greg Bowman,** PhD, Department of Biophysics, Johns Hopkins University, USA
gdbowman@jhu.edu, (410) 516-7305